

Review Article

Public Multi-Omics Databases for *In Silico* Cancer Research

Ahmed Mohammed Rahman¹ , Hawro Taha Hamza^{2,*} 

1 Department of Genetics, Bio Diagnostic Center (BDC), Erbil, 44001, Iraq

2 Department of Oncology, Nanakali Teaching Hospital, Erbil, 44001, Iraq

* Correspondence

hawrotaha@gmail.com

Article Info

Received: Mar 24, 2026

Revised: Apr 28, 2026

Accepted: May 04, 2026

Published: May 30, 2026

Abstract

Public multi-omics resources have made *in silico* cancer research a powerful approach for biomarker discovery, molecular stratification, pathway analysis, and therapeutic target prioritization. However, their value depends not only on data availability but also on the structured selection, retrieval, and integration of appropriate databases across multiple biological layers. This review provides an overview of the major public databases and analytical tools used in cancer data retrieval, interpretation, and translational inference. Genomics resources, including GenBank, Ensembl, gnomAD, dbSNP, COSMIC, ClinVar, dbVar, the GDC Portal, GeneCards, and the UCSC Genome Browser, support sequence retrieval, genome annotation, population filtering, somatic mutation analysis, and clinical variant interpretation. Transcriptomics databases, such as GEO, GTEx, TCGA, GENT2, SRA, Single Cell Expression Atlas, NONCODE, miRBase, Expression Atlas, FANTOM5, and BioGPS, enable tumor-normal comparison, expression profiling, regulatory RNA investigation, and cellular-resolution analysis. Proteomics databases extend interpretation to protein identity, abundance, interactions, and structural context, while metabolomics and pathway resources support biochemical and pathway-level analysis. DepMap adds a functional layer through cancer dependency and gene-essentiality data, strengthening target discovery and vulnerability assessment. Database Commons is highlighted as a useful meta-resource for database discovery. Collectively, these platforms support a stepwise framework for building evidence in computational oncology and translational decision-making.

Keywords: Cancer bioinformatics, *In Silico* oncology, Cancer data retrieval, Genomics, Transcriptomics, Proteomics, Metabolomics.



© 2026 Ahmed Mohammed Rahman and Hawro Taha Hamza. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license.

INTRODUCTION

Cancer is increasingly recognized as a systems-level disease driven by coordinated disturbances across multiple molecular layers, including genomic variation, transcriptional dysregulation, proteomic remodeling, and metabolic reprogramming¹. This molecular complexity underlies the marked heterogeneity observed among tumor types, between patients with the same histopathologic diagnosis, and even among cellular subpopulations within an individual tumor. As a result, contemporary oncology has

moved beyond single-marker investigation toward multi-omics research, in which disease mechanisms are explored through integrated evidence derived from DNA, RNA, proteins, metabolites, and functionally relevant cellular dependencies². Large international efforts have accelerated this transition by generating standardized cancer datasets across multiple platforms and by making them publicly available for secondary analysis, thereby establishing *in silico* investigation as a central component of modern cancer research.

In parallel with the expansion of high-

throughput sequencing, expression profiling, mass spectrometry, and pathway-mapping technologies, biological databases have become a fundamental infrastructure for discovery³. These resources no longer function merely as storage repositories; instead, they serve as analytical entry points through which researchers retrieve, compare, filter, and contextualize cancer-relevant molecular information⁴. However, the rapid growth in the number, specialization, and heterogeneity of public databases has introduced a new challenge. Investigators must now determine not only what type of data is needed, but also which resources are most appropriate for obtaining that evidence and how information from distinct repositories should be connected in a biologically meaningful way⁵. Thus, the central problem in *in silico* cancer research is no longer simply data availability, but the structured discovery, retrieval, and interpretation of data across multiple omics layers.

This challenge makes database discovery a critical first step in computational oncology. Before a researcher can assess somatic mutation recurrence, tumor-specific gene expression, protein abundance, pathway dysregulation, or therapeutic relevance, the broader database landscape must first be identified⁶. In this context, Database Commons (<https://ngdc.cncb.ac.cn/databasecommons/>) occupies an important methodological position. It serves as a curated catalog of biological databases that facilitates discovery of relevant resources according to category, data type, and biological scope, and is therefore especially useful at the gateway stage of *in silico* study design rather than within any single omics layer. Its role is

particularly valuable in cancer bioinformatics, where genomic, transcriptomic, proteomic, metabolomic, and clinical evidence are distributed across many specialized repositories⁷.

An additional development in this evolving landscape is the growing importance of functional dependency resources, particularly DepMap (<https://depmap.org/portal/>). Unlike databases that primarily describe molecular alterations or expression states, DepMap is designed to help researchers identify cancer vulnerabilities by providing open access to dependency data, analytical tools, and visualization resources across large panels of cancer cell lines. Its Achilles component systematically catalogs gene essentiality using genome-scale RNAi and CRISPR-Cas9 loss-of-function screens and links these dependencies to the molecular features of cancer cells, thereby supporting target prioritization and vulnerability mapping in precision oncology. In this sense, DepMap extends the logic of multi-omics investigation by adding a functional dimension that can help determine whether an altered gene or pathway is not only biologically relevant, but also selectively required for tumor survival⁸.

Against this background, the present review focuses on the practical and conceptual framework of *in silico* cancer data retrieval. It begins broadly by situating cancer as a multi-layered molecular disease and public databases as essential research infrastructure. It then narrows toward the methodological core of the topic, namely how major resources are organized across genomics, transcriptomics, proteomics, metabolomics, interaction biology, pathway analysis, and clinical or functional

interpretation, what each database contributes, and how these resources can be used in a stepwise retrieval and data-mining workflow. Finally, it widens again toward the broader implications of structured multi-omics and functional-data integration for biomarker discovery, mechanistic inference, target prioritization, and translational oncology. The database framework examined in this review is based on the current manuscript inventory, which collectively covers major repositories and analytical tools across the principal domains relevant to *in silico* cancer research. Within the genomic layer, resources such as GenBank, Ensembl, gnomAD, the 1000 Genomes Project, dbSNP, COSMIC, the GDC Portal, GeneCards, ClinVar, dbVar, and the UCSC Genome Browser support sequence retrieval, genome annotation, background variation filtering, somatic mutation interpretation, and clinical or structural variant assessment. Transcriptomic resources, including GEO, GTEx, TCGA, GENT2, SRA, Single Cell Expression Atlas, NONCODE, miRBase, Expression Atlas, FANTOM5, and BioGPS, extend this framework toward bulk expression analysis, healthy-tissue comparison, raw RNA-seq reanalysis, regulatory RNA exploration, and cellular heterogeneity. Proteomics resources such as UniProt, Human Protein Atlas, PRIDE, MassIVE, PeptideAtlas, STRING, PDB, ProteomicsDB, neXtProt, PaxDB, and CORUM provide the means to examine protein identity, abundance, interaction networks, structural context, and experimental validation. Metabolomics databases, including KEGG, HMDB, Reactome, PubChem, PathWhiz, MetaboLights, and SMPDB, widen biological interpretation by connecting molecular alterations to pathway-level and biochemical consequences. Clinical and translational

resources such as TCGA, ICGC, OncoKB, CIViC, cBioPortal, ClinicalTrials.gov, and SEER, together with functional platforms such as DepMap, further strengthen this ecosystem by linking molecular findings to therapeutic evidence, clinical outcomes, and cancer dependencies. Figure 1 provides an overview of the major public database categories used in *in silico* cancer research.

The main objective of this review is therefore not simply to enumerate databases, but to reframe them as components of a coherent evidence-building pathway. In a robust *in silico* cancer workflow, genomic repositories define alteration, transcriptomic repositories evaluate activity, proteomic repositories assess functional embodiment, metabolomic and pathway repositories reveal systems-level consequence, and clinical and functional resources help determine translational and therapeutic relevance. By describing these resources within a unified analytical structure, this review aims to clarify how database discovery, data retrieval, cross-omics interpretation, and dependency-informed prioritization can be organized into a reproducible and biologically meaningful strategy for computational oncology.

DATABASE DISCOVERY BEFORE DATABASE RETRIEVAL

A critical but frequently underexplained stage in *in silico* cancer research is database discovery. Before investigators retrieve any sequence, expression profile, protein record, or metabolite map, they must first determine which database type is appropriate for the biological question under investigation. This step is methodologically important because cancer bioinformatics operates across multiple data layers, each of which is supported by repositories with different scopes, structures, and intended uses⁹. A

resource designed for population-scale variant frequency, for example, does not serve the same purpose as a tumor-centered mutation archive, an RNA expression repository, a proteomics experiment database, or a

pathway knowledgebase¹⁰. Therefore, rigorous *in silico* study design begins not with downloading data, but with identifying the correct database class and then selecting the most suitable repository within that class.

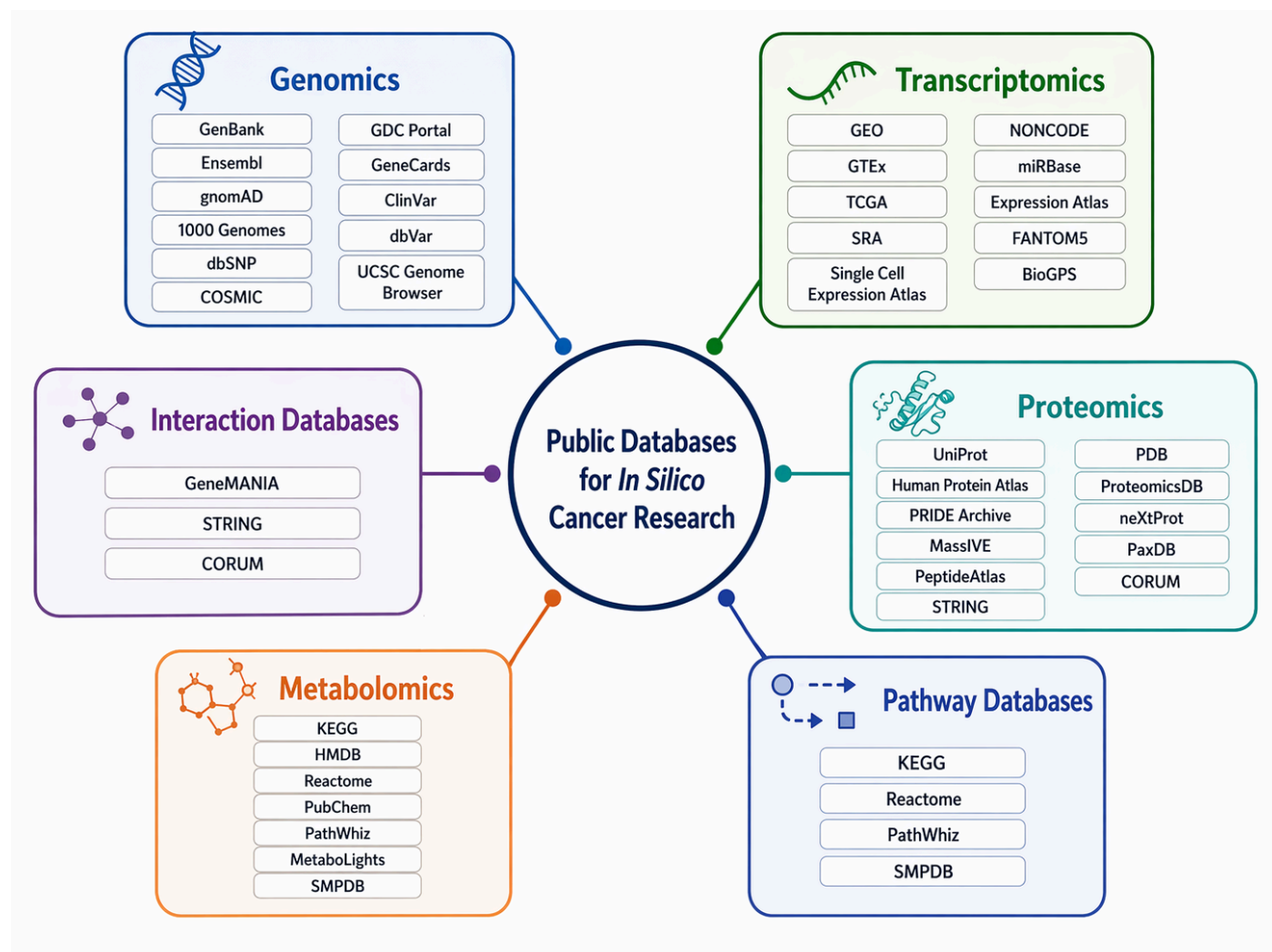


Figure 1. Public database landscape for in silico cancer research. Major resources are grouped into genomics, transcriptomic, proteomic, and metabolomic analysis, and concludes with cross-omics how different database classes contribute complementary layers of molecular information.

Within this context, Database Commons occupies a distinct and valuable role. It should not be regarded as a primary omics repository in the same sense as GenBank, GEO, PRIDE, or KEGG. Rather, Database Commons is a curated catalog of worldwide biological databases that was developed to provide a broad landscape of biological data resources and to enable users to discover databases according to factors such as database name, category, data type, species, institution, and subject area. The official Database Commons

resource describes itself as a manually curated and regularly enriched catalog, while its companion publication emphasizes that it was created to organize biological databases globally and make their associated meta-information easier to retrieve and compare. This makes Database Commons especially useful at the pre-retrieval stage of cancer bioinformatics, when the investigator is still mapping the database ecosystem and deciding where the most relevant evidence is likely to be found⁷.

The relevance of database discovery is particularly evident in oncology because cancer-related evidence is inherently distributed across multiple molecular domains. A researcher studying a candidate cancer gene may need one resource to define genomic coordinates and transcript models, another to assess allele frequency in reference populations, another to examine somatic mutation recurrence in tumors, another to compare tumor and normal expression, and additional repositories to evaluate protein abundance, interaction networks, or metabolic pathway consequences¹¹. Without an explicit discovery step, this process becomes fragmented and potentially biased, as investigators may default to familiar resources while overlooking more appropriate or higher-quality databases. Database Commons helps reduce that risk by functioning as a meta-resource that supports rational repository selection before primary data mining begins¹².

From a workflow perspective, database discovery should therefore be treated as the gateway between the initial research question and the subsequent stages of data retrieval, quality filtering, and integrative analysis. In a well-designed *in silico* cancer study, the sequence begins with clarification of the biological problem, proceeds to database discovery and classification, and only then advances toward omics-specific dataset acquisition¹³. In this review, Database Commons is positioned precisely at that gateway stage. It provides the conceptual bridge between the broad public data landscape and the narrower operational phase of database querying, thereby fitting naturally within the hourglass model of writing and study design. After this discovery phase, investigators can move more systematically

into genomics, transcriptomics, proteomics, and metabolomics repositories with a clearer rationale and a more reproducible retrieval strategy.

GENOMICS DATABASES in CANCER DATA RETRIEVAL

Genomics Databases

Genomics represents the foundational layer of *in silico* cancer research because many oncogenic processes originate from DNA-level abnormalities, including single-nucleotide substitutions, small insertions and deletions, copy-number changes, and larger structural rearrangements. For this reason, genomic data retrieval is usually the first analytical stage in computational oncology. However, genomic investigation is not supported by a single universal repository¹⁴. Instead, it depends on a set of complementary databases, each serving a distinct function in sequence retrieval, genome annotation, population filtering, somatic mutation interpretation, clinical classification, structural-variant assessment, or locus visualization. A rigorous cancer bioinformatics workflow should therefore treat these databases not as interchangeable resources, but as sequential components of an evidence-building system¹⁵.

GenBank serves as the primary entry point for sequence-centered investigations. As the largest public repository of DNA sequences submitted by research laboratories worldwide, it provides direct access to nucleotide records that are essential for basic sequence retrieval, reference comparison, accession-based confirmation, and preliminary molecular characterization. In cancer studies, GenBank is especially useful when investigators need canonical sequence records for a target gene, genomic fragment, or transcript before moving into downstream

annotation or mutation analysis. It functions best as a foundational archive: it defines what the sequence is, where it can be retrieved, and which molecular record is being analyzed¹⁶.

Ensembl builds upon this sequence foundation by supplying a far richer annotation environment. Whereas GenBank primarily supports sequence access, Ensembl provides genome annotation, gene models, transcript structures, comparative genomics, and visualization tools across multiple species. In practical cancer bioinformatics, Ensembl is critical when researchers need to determine exon–intron organization, transcript isoforms, chromosomal location, or conserved genomic features. It is also particularly valuable for aligning a candidate gene with its broader genomic architecture, thereby allowing mutations or expression changes to be interpreted in the context of known biological structure¹⁷.

gnomAD is central to the population-filtering step of genomic interpretation. Its major value lies in aggregating large-scale human genetic variation data and enabling the study of allele frequencies across populations. This function is indispensable in cancer bioinformatics because a candidate variant observed in a tumor cannot be interpreted meaningfully unless it is distinguished from common background variation. A variant that occurs frequently in the general population is less likely to represent a high-priority cancer driver, whereas a rare variant or one absent from reference populations may warrant deeper functional or clinical investigation. Thus, gnomAD acts as a critical filter between raw variant observation and cancer-specific interpretation¹⁸.

The 1000 Genomes Project complements gnomAD by providing a

reference dataset of human genetic variation derived from diverse global populations. Its importance lies in the breadth of population representation and its continued value as a comparative framework for human polymorphism. In the context of cancer studies, this resource helps researchers evaluate whether a detected alteration reflects common human diversity or may instead represent a rarer genomic event with stronger disease relevance. While gnomAD is often used for large-scale allele-frequency filtering, the 1000 Genomes Project provides an additional population-genetic reference that strengthens background comparison and reduces the risk of overinterpreting naturally occurring variants as tumor-specific findings¹⁹.

dbSNP remains one of the most important repositories for single nucleotide polymorphisms and other small genetic variants. Its particular importance lies in standardization, identifier tracking, and broad catalog coverage. In cancer data retrieval, dbSNP is useful when investigators need to determine whether a variant has been previously registered, whether a recognized identifier exists for that change, and whether the observed variant belongs to the wider universe of known human small-scale genomic variation. Although dbSNP does not by itself establish oncogenic importance, it is an essential reference layer in the chain of variant characterization²⁰.

COSMIC occupies a uniquely important role in cancer genomics because it is explicitly focused on somatic mutations in human cancer. Unlike population repositories, COSMIC is designed to answer whether a mutation, gene, or genomic region has already been implicated in tumor biology. Its usefulness lies in recurrence mapping, cancer-

type association, and somatic mutation cataloging. In a professional *in silico* cancer workflow, COSMIC is often the point at which a filtered variant begins to acquire true oncologic context. If a rare alteration also appears recurrently in tumor datasets cataloged by COSMIC, its potential biological significance increases substantially²¹.

The GDC Portal extends cancer genomics beyond mutation listing by providing access to integrated cancer genomic datasets that include sequencing data as well as clinical information. This makes it particularly valuable for studies that do not stop at variant discovery but aim to connect genomic abnormalities to patient-level or tumor-level features. In practice, the GDC Portal is one of the most important translational resources in the genomics layer because it links molecular data to clinical context, thereby allowing investigators to move from descriptive mutation mining toward clinically informed cancer interpretation²².

GeneCards serves a different but equally useful purpose. Rather than functioning as a raw-data repository, it is a comprehensive gene-centered knowledgebase that brings together gene function, expression, and disease associations. Its strength lies in integration. When a cancer study begins with a candidate gene rather than a specific variant, GeneCards helps the researcher quickly assemble a functional profile of that gene, including its biological roles, related diseases, and expression context. In this sense, GeneCards often acts as the conceptual bridge between sequence-level information and broader biological interpretation²³.

ClinVar is essential when variant interpretation moves into the clinical domain. It contains information on the relationship

between genetic variants and their clinical significance, making it especially relevant when a mutation is suspected to have diagnostic, prognostic, or pathogenic value. In cancer data retrieval, ClinVar helps determine whether a variant has already been clinically evaluated and whether any interpretive consensus exists around its medical relevance. Although not every tumor-associated variant is represented in ClinVar, its presence can significantly strengthen the translational relevance of a candidate genomic alteration²⁴.

dbVar adds the large-scale structural dimension to cancer genomics. Many tumors are not defined only by small variants, but also by deletions, duplications, and other structural rearrangements that can profoundly alter gene dosage and genomic architecture. dbVar is important because it specifically catalogs these large-scale genomic changes. In a comprehensive cancer bioinformatics workflow, it helps ensure that the investigator does not restrict genomic interpretation only to SNPs and indels, but also considers structural events that may be central to tumor biology²⁵.

Finally, the UCSC Genome Browser provides the interactive visualization layer that ties the genomic framework together. Its major importance lies in allowing researchers to explore genomic loci with annotations, comparative information, and integrated tracks. Rather than viewing a candidate alteration as a single isolated data point, investigators can use the UCSC Genome Browser to examine the surrounding gene context, structural features, and relevant annotations. This makes it a critical interpretive endpoint in genomic data retrieval, where results are contextualized within broader genome²⁶. The

major genomics databases used in *in silico* cancer research are summarized in Table 1.

Table 1. Genomics databases for in silico cancer data retrieval

Database	Description	Link
GenBank	Largest public database of DNA sequences submitted from worldwide research laboratories; used for sequence storage and retrieval.	https://www.ncbi.nlm.nih.gov/genbank/
Ensembl	Provides genome annotation, gene models, comparative genomics, and visualization tools for multiple species.	https://www.ensembl.org
gnomAD	Aggregates large-scale human genetic variation data to study population-level allele frequencies.	https://gnomad.broadinstitute.org
1000 Genomes Project	Provides a reference dataset of human genetic variation from diverse global populations.	https://www.internationalgenome.org
dbSNP	Repository of single nucleotide polymorphisms and small genetic variations.	https://www.ncbi.nlm.nih.gov/snp/
COSMIC	Catalog of somatic mutations in human cancer.	https://cancer.sanger.ac.uk/cosmic
GDC Portal	Provides cancer genomic datasets including sequencing and clinical data.	https://portal.gdc.cancer.gov
GeneCards	Comprehensive database providing gene function, expression, and disease associations.	https://www.genecards.org
ClinVar	Contains relationships between genetic variants and clinical significance.	https://www.ncbi.nlm.nih.gov/clinvar/
dbVar	Database for large-scale structural genomic variations such as deletions and duplications.	https://www.ncbi.nlm.nih.gov/dbvar/
UCSC Genome Browser	Interactive platform for exploring genomes with annotations and comparative data.	https://genome.ucsc.edu

Collectively, these databases constitute a genomics ecosystem that is structured and nonredundant, with the purpose of retrieving cancer data. GenBank and Ensembl define the sequence and genomic structure; gnomAD, the 1000 Genomes Project, and dbSNP filter and contextualize background variation; COSMIC and the GDC Portal establish cancer-specific relevance; GeneCards links genes to function and disease; ClinVar and dbVar refine clinical and structural interpretation; and the UCSC Genome Browser provides an annotated visualization platform. Used sequentially, they allow the researcher to move from a raw genomic query to a biologically and clinically grounded cancer interpretation.

Genomics Tools

Genomics tools are essential for transforming raw sequence data into interpretable evidence that can support cancer-related discovery and variant characterization. In this context, GeneMANIA is useful for gene-function prediction and gene–gene interaction analysis, allowing candidate genes to be placed within a broader functional and network-based context.²⁷ Basic Local Alignment Search Tool (BLAST) remains a foundational sequence-analysis tool because it identifies sequence similarity between DNA or protein sequences and is widely used for sequence confirmation, homology assessment, and reference

comparison.²⁸ MutaGene adds a more cancer-specific layer of interpretation by enabling analysis of somatic mutation patterns, mutational signatures, and the distinction between likely driver and passenger variants in tumor genomes. For read-mapping workflows, BWA and Bowtie2 are widely used to align short sequencing reads to a reference genome, thereby forming the basis for downstream genomic analysis.²⁹ GATK then provides a major variant-discovery framework,

particularly for SNP and indel detection, and is one of the most important tools in cancer genomics pipelines.³⁰ Finally, Integrative Genomics Viewer (IGV) functions as an interactive genomic visualization platform that allows researchers to inspect alignments, variants, and genomic features directly, which is particularly valuable for quality control and result interpretation.³¹ The core genomics analytical tools used for sequence analysis and variant interpretation are listed in Table 2.

Table 2. Genomics tools for cancer sequence analysis and variant interpretation

Tools	Description	Link
GeneMANIA	Gene function prediction and gene-gene interaction network analysis tool	https://genemania.org
BLAST	Identifies sequence similarity between DNA/protein sequences using alignment search	https://blast.ncbi.nlm.nih.gov/Blast.cgi
MutaGene	Web-based bioinformatics platform used to analyze cancer mutational signatures, study somatic mutation patterns, and distinguish driver vs passenger mutations in tumor genomes	https://www.ncbi.nlm.nih.gov/research
BWA	Aligns short DNA sequencing reads to a reference genome	https://bio-bwa.sourceforge.net
GATK	Toolkit for variant discovery (SNPs and indels) from sequencing data	https://gatk.broadinstitute.org
Bowtie2	Fast and memory-efficient alignment of sequencing reads to reference genomes	http://bowtie-bio.sourceforge.net/bowtie2
IGV	Interactive tool for visualization of genomic data such as alignments and variants	https://igv.org

TRANSCRIPTOMICS DATABASES and TOOLS in CANCER DATA RETRIEVAL
Transcriptomics Databases

Transcriptomics occupies a central position in *in silico* cancer research because it provides direct insight into the functional consequences of genomic change. While genomics identifies the underlying molecular substrate of cancer, transcriptomics reveals which genes, transcripts, and regulatory RNAs are actively expressed, suppressed, or altered across tumor states.³² In cancer data retrieval, transcriptomic databases are therefore indispensable for defining tumor-associated

expression signatures, comparing malignant and non-malignant tissues, investigating regulatory RNA biology, and exploring cellular heterogeneity within the tumor microenvironment. The transcriptomics resources in your attached framework are not redundant; rather, each database contributes a distinct layer of evidence and should be used according to its specific strength within the analytical workflow.³³

Gene Expression Omnibus (GEO) is one of the most important public repositories for transcriptomic studies because it stores gene-expression datasets generated from both

microarray and RNA-seq experiments. Its importance lies in breadth and reusability. GEO allows researchers to access published expression studies across many cancer types, platforms, and experimental designs, making it highly valuable for secondary analysis, comparative expression studies, and validation of candidate biomarkers in independent cohorts. In practical cancer bioinformatics, GEO is often one of the first resources consulted when the objective is to identify whether a gene or transcript is consistently dysregulated across multiple tumor studies³⁴.

Genotype-Tissue Expression (GTEx) is particularly important because it provides tissue-specific gene-expression profiles in healthy human tissues. Its value in cancer transcriptomics lies in its function as a normal reference framework. Tumor expression data alone may show that a transcript is highly abundant, but such abundance cannot be interpreted as abnormal unless it is compared against healthy tissue baselines. GTEx therefore plays a critical comparative role by helping investigators distinguish genuine tumor-associated dysregulation from physiologic tissue-specific expression. In many cancer studies, GTEx is essential for adding biological context and for preventing overinterpretation of tumor-only observations³⁵.

The Cancer Genome Atlas (TCGA) is one of the most influential resources in transcriptomic oncology because it provides large-scale cancer genomics and transcriptomics datasets that are linked to clinical and molecular annotations. Its particular importance lies in the depth of cancer-specific coverage. Unlike broader public repositories, TCGA is specifically

structured around tumor cohorts, which makes it exceptionally useful for identifying cancer-associated expression patterns, comparing subtypes, relating transcripts to clinical outcome, and integrating RNA data with genomic alterations. In cancer data retrieval, TCGA often serves as the principal source for tumor-centered transcriptomic discovery³⁶.

GENT2 is a cancer-focused gene expression database that allows users to compare gene expression between normal and tumor tissues using uniformly processed public microarray datasets. It is especially useful for validating candidate genes, examining tissue- and subtype-specific expression patterns, and exploring possible clinical relevance through integrated survival analysis features. For *in silico* cancer research, GENT2 provides a convenient bridge between large public expression datasets and practical biological interpretation³⁷.

Sequence Read Archive (SRA) is crucial because it stores raw high-throughput sequencing data, including RNA-seq datasets. Its importance differs from that of GEO or TCGA: rather than functioning primarily as a curated expression repository, SRA supports reanalysis from raw data. This makes it particularly valuable for researchers who need to apply standardized pipelines, perform custom alignment and quantification, or verify published findings independently. In transcriptomic cancer research, SRA provides methodological flexibility and supports reproducibility by allowing investigators to begin from the original sequencing reads rather than relying only on processed outputs³⁸.

Single Cell Expression Atlas has become increasingly important in cancer research

because it provides single-cell RNA sequencing data across multiple species and conditions. Its main contribution is cellular resolution. Bulk transcriptomic datasets average signals across many cells, which can obscure tumor heterogeneity, immune infiltration, stromal contributions, and rare malignant subpopulations. By contrast, single-cell datasets allow investigators to ask which specific cell populations express a transcript of interest and how transcriptional states vary within the tumor ecosystem. This makes Single Cell Expression Atlas particularly valuable for modern cancer studies focused on microenvironmental biology, lineage tracing, and intratumoral diversity³⁹.

NONCODE is an important database because it specializes in non-coding RNA sequences and annotations. Its significance lies in expanding transcriptomic analysis beyond protein-coding genes. In cancer biology, long non-coding RNAs and other non-coding transcripts are increasingly recognized as regulators of proliferation, invasion, metastasis, immune escape, and therapy response. NONCODE therefore provides a dedicated resource for identifying and contextualizing non-coding RNA candidates that may be missed in more gene-centered or protein-oriented repositories. In transcriptomic cancer retrieval, it is especially useful when the study objective includes regulatory RNA mechanisms rather than only coding-gene expression⁴⁰.

miRBase is the principal repository in this framework for microRNA sequences and annotation. Its importance lies in the fact that microRNAs are major post-transcriptional regulators of gene expression and play well-established roles in tumor initiation, progression, metastasis, and resistance. In *in*

silico cancer data retrieval, miRBase is essential for identifying microRNA entities accurately, standardizing nomenclature, and connecting sequence information to downstream functional interpretation. When cancer studies investigate regulatory networks or non-coding RNA biomarkers, miRBase provides a critical reference layer⁴¹.

Expression Atlas provides gene and transcript expression data across tissues, species, and conditions, and its value lies in broad comparative context. Whereas TCGA is strongly cancer-centered and GTEx is focused on healthy tissue reference, Expression Atlas allows investigators to place a transcript within a wider expression landscape. This is useful when evaluating whether a gene is condition-specific, broadly expressed, or selectively altered in particular biological states. In cancer transcriptomics, Expression Atlas strengthens interpretation by helping researchers compare tumor-associated signals against a wider universe of biological expression profiles³⁹.

FANTOM5 is especially important because it extends transcriptomic analysis beyond conventional gene expression and incorporates functional annotation of mammalian genomes, including promoter and enhancer RNA expression. Its significance lies in regulatory interpretation. Cancer is often driven not only by changes in transcript abundance, but also by altered promoter usage, enhancer activation, and transcriptional control landscapes. FANTOM5 therefore becomes particularly valuable in studies that seek to understand transcriptional regulation at a deeper mechanistic level, especially when the research question involves promoter activity, enhancer-associated transcription, or non-canonical regulatory elements⁴².

Table 3. Transcriptomics databases for in silico cancer data retrieval

Database	Description	Link
GEO (Gene Expression Omnibus)	Public repository of gene expression data from microarray and RNA-seq experiments	https://www.ncbi.nlm.nih.gov/geo/
GTEx (Genotype-Tissue Expression)	Provides tissue-specific gene expression profiles in humans	https://gtexportal.org
TCGA (The Cancer Genome Atlas)	Large-scale cancer genomics and transcriptomics dataset	https://www.cancer.gov/ccg/research/genome-sequencing/tcga
GENT2	A platform for exploring Gene Expression patterns across Normal and Tumor tissues	http://gent2.appex.kr/
SRA (Sequence Read Archive)	Raw high-throughput sequencing data including RNA-seq datasets	https://www.ncbi.nlm.nih.gov/sra
Single Cell Expression Atlas	Single-cell RNA sequencing data across multiple species and conditions	https://www.ebi.ac.uk/gxa/sc/home
NONCODE	Database of non-coding RNA sequences and annotations	http://www.noncode.org
miRBase	Repository for microRNA sequences and annotation	https://www.mirbase.org
Expression Atlas	Provides gene and transcript expression across species, tissues, and conditions	https://www.ebi.ac.uk/gxa/home
FANTOM5	Functional annotation of mammalian genome including promoter and enhancer RNA expression	https://fantom.gsc.riken.jp/5/
BioGPS	Gene annotation and expression portal for multiple organisms	http://biogps.org

BioGPS serves as a gene annotation and expression portal across multiple organisms and contributes a flexible exploratory dimension to transcriptomic data retrieval. Its importance lies in accessibility and breadth. BioGPS allows rapid interrogation of gene-expression patterns across different tissues and systems, which is useful when researchers want an initial overview of where and how strongly a gene is expressed before moving into more specialized cancer-focused analyses. In this sense, BioGPS often functions as a supportive exploratory tool that complements more disease-specific repositories such as TCGA and GEO⁴³. The principal transcriptomics databases used in cancer data retrieval are presented in Table 3.

Taken together, these transcriptomics

databases form a layered and highly complementary framework for cancer data retrieval. GEO and TCGA provide major tumor-expression resources; GTEx supplies healthy tissue context; SRA enables raw-data reanalysis; Single Cell Expression Atlas resolves cellular heterogeneity; NONCODE and miRBase expand the analysis into regulatory RNA biology; Expression Atlas and BioGPS add broad comparative context; and FANTOM5 deepens regulatory interpretation. Used sequentially and critically, these databases allow investigators to move from general transcript identification to biologically and clinically meaningful expression analysis in cancer. Figure 2 shows how genomic findings can be progressively strengthened through transcriptomic validation in cancer studies.

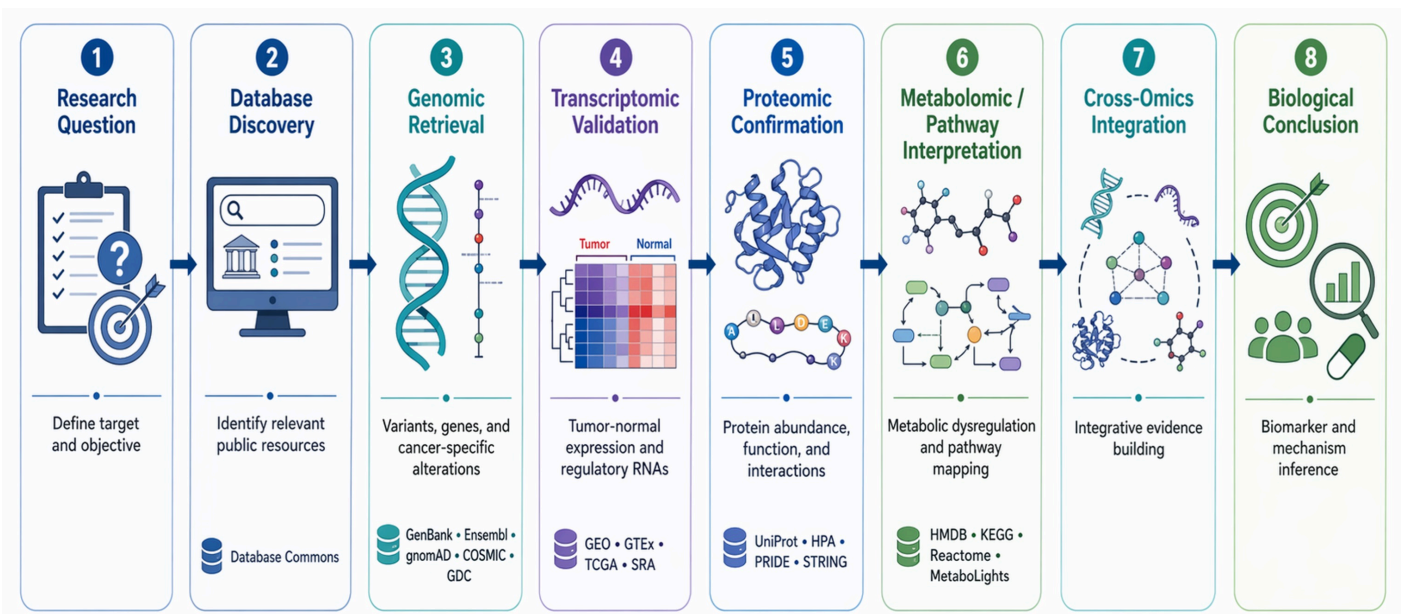


Figure 2. Genomics-to-transcriptomics evidence-building framework. The workflow progresses from gene and sequence definition to population filtering, cancer-specific genomic interpretation, and transcriptomic validation using complementary public resources.

Transcriptomics Tools

Transcriptomics tools are central to the analysis of RNA-derived data in cancer research because they enable the identification of transcriptional changes associated with tumor biology. STAR is a high-performance RNA-seq aligner that supports splice-junction detection and is therefore especially well suited for transcriptome mapping⁴⁴. StringTie complements alignment by assembling transcripts and estimating transcript abundance, making it highly useful for transcript-level characterization and downstream expression analysis.⁴⁵ For statistical analysis of gene-expression

differences, Limma, edgeR, and DESeq2 represent the core analytical tools in many cancer transcriptomics studies. Limma is particularly valuable for microarray data and adapted RNA-seq workflows,⁴⁶ while edgeR and DESeq2 are specifically optimized for count-based RNA-seq data and provide robust models for detecting differentially expressed genes.⁴⁷ Together, these tools form an integrated analytical framework for alignment, quantification, and statistical evaluation of cancer-associated transcriptomic changes. The main transcriptomics tools for cancer expression analysis are summarized in Table 4.

Table 4. Transcriptomics tools for cancer expression analysis

Tools	Description	Link
STAR	Ultra-fast alignment of RNA-seq reads to reference genome with splice-junction detection	https://github.com/alexdobin/STAR
StringTie	Assembles transcripts and estimates their abundance from RNA-seq data	https://ccb.jhu.edu/software/stringtie/
Limma	Statistical tool for differential expression analysis (microarray and RNA-seq)	https://bioconductor.org/packages/release/bioc/html/limma.html
edgeR	Performs differential gene expression analysis using count-based RNA-seq data	https://bioconductor.org/packages/release/bioc/html/edgeR.html
DESeq2	Identifies differentially expressed genes using negative binomial statistical models	https://bioconductor.org/packages/release/bioc/html/DESeq2.html

PROTEOMICS DATABASES and TOOLS in CANCER DATA RETRIEVAL

Proteomics Databases

Proteomics provides a critical interpretive layer in *in silico* cancer research because it connects upstream molecular abnormalities to their functional biological products. Although genomic and transcriptomic analyses can identify candidate genes and dysregulated transcripts, these alterations do not always translate directly into corresponding protein abundance, localization, interaction behavior, or structural consequence⁴⁸. For this reason, proteomics databases are indispensable in cancer data retrieval, particularly when the objective is to validate biomarker candidates, investigate tumor-associated protein expression, explore interaction networks, or understand structural and functional mechanisms at the protein level. The proteomics resources listed in your attached framework are highly complementary, and each database contributes a distinct form of evidence that strengthens cancer-related interpretation⁴⁹.

UniProt is the principal foundational database in the proteomics layer. Its importance lies in the breadth and depth of protein-centered information that it provides, including protein sequence, molecular function, domain architecture, catalytic activity, subcellular localization, and curated biological annotation. In cancer data retrieval, UniProt is often the first resource used to define the protein product of a candidate gene, clarify isoforms, and interpret its known functional roles. It is especially valuable when researchers need a reliable reference point before moving toward tissue expression, pathway integration, or protein interaction analysis. In this sense, UniProt serves as the core descriptive and functional

knowledgebase for protein identity in computational oncology⁵⁰.

Human Protein Atlas is particularly important because it extends protein analysis into tissue and cellular context. Its main value lies in mapping protein expression across human tissues, cells, and organs, thereby allowing researchers to determine where a protein is normally present and whether its distribution appears altered in disease settings. In cancer studies, this database is especially useful for evaluating tissue-specific protein abundance, examining tumor-associated overexpression, and comparing normal and malignant protein patterns. While UniProt defines what a protein is, Human Protein Atlas helps clarify where that protein is biologically relevant. This makes it one of the most useful translational bridges between computational findings and tissue-level cancer biology⁵¹.

The PRoteomics IDentifications (PRIDE) database Archive is a major repository for public mass spectrometry-based proteomics data and is particularly important for experimental validation and dataset reuse. In *in silico* cancer research, its value lies in providing access to previously generated proteomics experiments that can be reanalyzed or mined for protein-level confirmation. If a transcriptomic analysis suggests that a candidate gene is upregulated in tumors, PRIDE can be used to determine whether corresponding protein evidence is available in public mass spectrometry datasets. Thus, PRIDE is highly relevant when investigators want to move beyond knowledge-level annotation and toward experimental proteomic support⁵².

MassIVE plays a similar role to PRIDE, but its strength lies in serving as a large-scale

repository for proteomics mass spectrometry data, often supporting deep experimental archives and large datasets. Its importance in cancer bioinformatics lies in data accessibility and scalability. For studies that aim to validate proteomic signatures across multiple datasets or perform broad secondary analyses of public mass spectrometry experiments, MassIVE offers an important source of reusable evidence. It strengthens the proteomics layer by ensuring that experimental cancer proteomics data remain computationally accessible for independent interrogation.

PeptideAtlas is valuable because it focuses specifically on identified peptides and proteins derived from mass spectrometry experiments. Its importance lies in the peptide-level evidence it provides for protein existence and detectability. In cancer data retrieval, PeptideAtlas can help confirm whether the peptides corresponding to a protein of interest have been experimentally observed, which is especially useful in biomarker discovery and targeted proteomics planning. Compared with broader protein databases, PeptideAtlas offers a more experimentally anchored view of protein identification⁵³.

STRING is one of the most important resources for moving from individual proteins to functional network biology. Its principal value lies in protein-protein interaction analysis and the identification of functional associations. Cancer is rarely driven by isolated proteins acting alone; instead, tumor biology emerges through pathways, complexes, and interaction networks. STRING is therefore highly useful when a researcher needs to determine whether a candidate protein interacts with established oncogenic pathways, whether it occupies a central

position in signaling networks, or whether multiple cancer-associated proteins converge within a shared functional module. In proteomics-based cancer mining, STRING transforms protein lists into biologically interpretable systems⁵⁴.

Protein Data Bank (PDB) adds a structural dimension that is especially important in mechanistic cancer research. Its role lies to provide three-dimensional structural data for proteins and other biomolecules. In practical terms, PDB is essential when the significance of a cancer-associated mutation or protein alteration depends on its possible structural consequences, such as altered binding, folding instability, or disruption of catalytic sites. Structural evidence can greatly strengthen the plausibility of a candidate driver event, particularly when sequence-level or expression-level changes need mechanistic support. Thus, PDB moves proteomics interpretation from expression and annotation toward structural biology and molecular mechanism⁵⁵.

ProteomicsDB is important because it integrates human proteome expression and quantification data, thereby providing a quantitative view of protein abundance. In cancer data retrieval, this database is useful when investigators want to evaluate how strongly a protein is expressed across tissues or conditions and whether abundance patterns support tumor-associated interpretation. Its particular strength lies in linking proteomics evidence with broader expression trends, making it valuable for cross-validating protein abundance after transcriptomic screening⁵⁶.

neXtProt serves as a high-quality human protein knowledgebase with strong functional

annotation, and its importance lies in its curated human-protein focus. While UniProt is broad and foundational, neXtProt offers a more specialized human-centered interpretive layer that can be particularly useful in cancer studies, where the relevance of a protein often depends on detailed functional annotation within human biology. It helps refine understanding of protein role, evidence status, and biological significance in a way that complements broader annotation resources⁵⁷.

PaxDB is valuable because it provides protein abundance data across organisms and

tissues. Its main importance in cancer bioinformatics lies in contextual abundance assessment. When a researcher wants to know whether a protein is generally abundant, tissue-restricted, or relatively rare under physiologic conditions, PaxDB offers a broad abundance framework that can help contextualize tumor-associated observations. In this way, PaxDB contributes a comparative quantitative perspective that supports the interpretation of whether a cancer-related protein signal is truly exceptional or consistent with broader biologic abundance patterns⁵⁸.

Table 5. Proteomics databases for in silico cancer data retrieval

Database	Description	Link
UniProt	Comprehensive protein sequence and functional information database (Swiss-Prot & TrEMBL)	https://www.uniprot.org
Human Protein Atlas	Maps protein expression across human tissues, cells, and organs	https://www.proteinatlas.org
PRIDE Archive	Public repository for mass spectrometry-based proteomics data	https://www.ebi.ac.uk/pride/
MassIVE	Large-scale mass spectrometry proteomics data repository	https://massive.ucsd.edu
PeptideAtlas	Repository of identified peptides and proteins from MS experiments	http://www.peptideatlas.org
STRING	Protein–protein interaction networks and functional associations	https://string-db.org
PDB (Protein Data Bank)	3D structural data of proteins and biomolecules	https://www.rcsb.org
ProteomicsDB	Integrated human proteome expression and quantification data	https://www.proteomicsdb.org
neXtProt	High-quality human protein knowledgebase with functional annotation	https://www.nextprot.org
PaxDB	Protein abundance database across organisms and tissues	https://pax-db.org
CORUM	Database of experimentally verified protein complexes	https://mips.helmholtz-muenchen.de/corum/

CORUM occupies a distinct position in the proteomics landscape because it is focused on experimentally verified protein complexes. Its importance lies in revealing whether a protein of interest acts as part of a known functional complex rather than as an isolated molecular entity. This is especially relevant in cancer, where disruption of complex assembly, signaling modules, or

multiprotein regulatory systems may drive malignant behavior. In an *in silico* cancer workflow, CORUM helps determine whether a candidate protein should be interpreted in the context of a broader molecular machine, thereby deepening functional inference beyond single-protein analysis⁵⁹. The major proteomics databases used in *in silico* cancer research are listed in Table 5.

These proteomics databases provide a highly structured framework for cancer data retrieval. UniProt and neXtProt define protein identity and function; Human Protein Atlas, ProteomicsDB, and PaxDB clarify protein abundance and tissue relevance; PRIDE Archive, MassIVE, and PeptideAtlas provide experimental proteomic evidence; STRING and CORUM explain interaction and complex biology; and PDB contributes structural interpretation. Used sequentially and critically,

these resources allow researchers to move from a candidate gene or transcript to a biologically grounded protein-level understanding that is far more relevant to cancer mechanism and translational application. Figure 3 illustrates how a candidate protein can be interpreted through annotation, abundance profiling, experimental support, interaction analysis, and structural context in *in silico* cancer research.

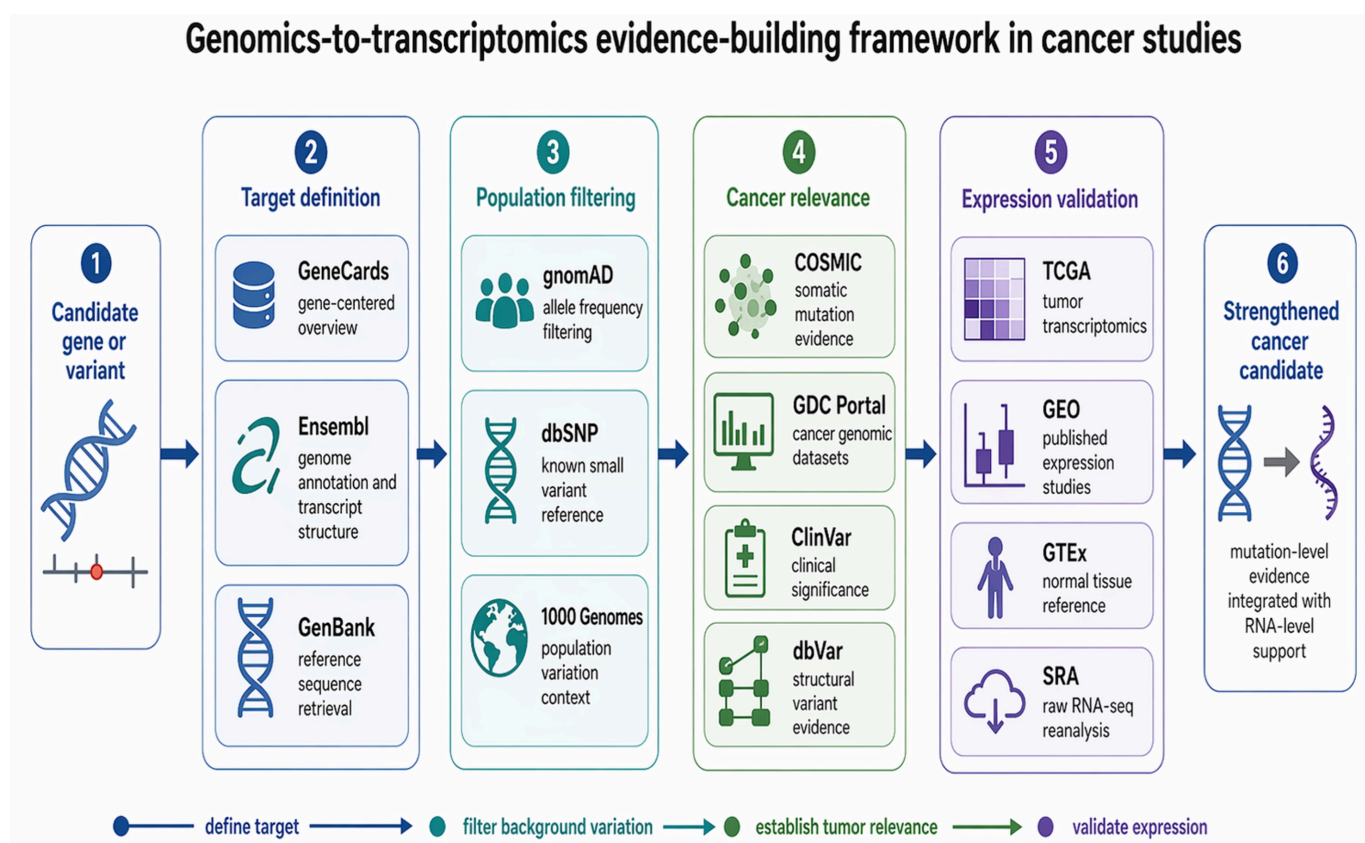


Figure 3. Proteomic, interaction, and structural interpretation framework. Candidate proteins are evaluated through functional annotation, tissue abundance, experimental proteomics evidence, interaction networks, protein complexes, and three-dimensional structural information.

Proteomics Tools

Proteomics tools play a critical role in determining whether genomic or transcriptomic findings are reflected at the protein level and in quantifying proteome-wide changes in cancer. MaxQuant is one of the most widely used proteomics platforms because it supports both protein identification and quantitative analysis from mass spectrometry data. Its importance lies in converting raw proteomics experiments into

interpretable protein abundance and identification outputs⁶⁰. Skyline is especially valuable for targeted proteomics and quantitative mass spectrometry analysis, and it is often used when researchers need focused validation of peptides or proteins of interest⁶¹. ProteinPilot also supports protein identification from MS/MS data and contributes to peptide-spectrum matching and proteome characterization. Collectively, these

tools provide the computational foundation for transforming experimental mass spectrometry data into meaningful protein-level evidence in cancer bioinformatics⁶². The core proteomics analytical tools used for protein identification and quantitative analysis are shown in Table 6.

Table 6. Proteomics tools for protein identification and quantitative analysis

Tool	Description	Link
MaxQuant	Identifies and quantifies proteins from mass spectrometry (MS) data	https://www.maxquant.org
Skyline	Tool for targeted proteomics and quantitative MS data analysis	https://skyline.ms
ProteinPilot	Software for protein identification using MS/MS data	https://sciex.com/products/software/proteinpilot-software

METABOLOMICS DATABASES and TOOLS in CANCER DATA RETRIEVAL
Metabolomics Databases

Metabolomics occupies a particularly important position in *in silico* cancer research because it captures the biochemical consequences of upstream genomic, transcriptomic, and proteomic alterations. While DNA- and RNA-level analyses identify potential drivers of malignant behavior, metabolomics helps determine how these changes are translated into altered pathway activity, substrate utilization, redox balance, lipid remodeling, and other hallmarks of tumor metabolism⁶³. In cancer data retrieval, metabolomics databases are therefore essential not only for identifying metabolites and pathways, but also for connecting molecular abnormalities to functional cellular states. The metabolomics resources listed in your attached framework are highly complementary, and each database contributes a distinct type of information that strengthens pathway-level interpretation in computational oncology⁶⁴.

Kyoto Encyclopedia of Genes and Genomes (KEGG) is one of the most influential pathway-oriented resources in metabolomics and cancer bioinformatics. Its importance lies in its ability to connect genes, enzymes, metabolites, and biochemical reactions within integrated pathway maps. In practical cancer

research, KEGG is especially valuable when investigators want to place a dysregulated gene, enzyme, or metabolite into a broader metabolic or signaling context. It allows researchers to move from isolated molecular observations toward a network-level understanding of cancer-related pathway disruption. For this reason, KEGG is often one of the first resources used when attempting to interpret whether a molecular alteration affects glycolysis, amino acid metabolism, lipid metabolism, oxidative phosphorylation, or other cancer-relevant pathways⁶⁵.

Human Metabolome Database (HMDB) is particularly important because it focuses on detailed information about human metabolites and their biological roles. Its strength lies in compound-level annotation. Whereas KEGG is primarily pathway-centered, HMDB is metabolite-centered and provides rich information on metabolite identity, chemical properties, biological relevance, and human-specific context. In *in silico* cancer data retrieval, HMDB is highly useful when researchers need to characterize a specific metabolite, clarify its physiologic role, or understand its relevance to human disease processes. It is especially valuable in studies that move from broad pathway mapping toward detailed interpretation of individual tumor-associated metabolites⁶⁶.

Reactome is another major pathway database, but its distinct importance lies in its highly curated representation of biological pathways, including both metabolism and signaling. This makes Reactome especially valuable in cancer studies where metabolite changes must be interpreted in relation not only to metabolic circuits but also to wider regulatory networks. For example, many cancer-associated metabolic alterations are tightly linked to growth signaling, inflammatory signaling, or stress-response pathways. Reactome is therefore particularly useful when the aim is to connect metabolic observations to broader systems biology rather than to metabolism alone. In this sense, it serves as a bridge between metabolomics and multi-pathway integrative oncology⁶⁷.

PubChem contributes an important chemical dimension to cancer metabolomics. Its main role is to provide compound structures, physicochemical properties, identifiers, and related chemical information. In a cancer bioinformatics workflow, PubChem is especially useful when metabolite interpretation requires chemical clarification, structure-based comparison, or confirmation of compound identity. Although it is not a pathway database in the same sense as KEGG or Reactome, its relevance is substantial because accurate metabolic interpretation often depends on knowing the structural and chemical features of the molecules under study. Thus, PubChem supports the precise identification and characterization of compounds within metabolomics-driven cancer analyses⁶⁸.

PathWhiz is valuable because it is designed for the visualization and construction of metabolic pathways. Its importance lies in representation and

communication. In many *in silico* cancer studies, biological interpretation becomes much stronger when complex metabolic results can be organized visually into pathway models. PathWhiz supports this need by enabling researchers to create or explore pathway diagrams that make biochemical relationships more explicit. In a review or integrative study, it is particularly useful for translating lists of altered metabolites or enzymes into interpretable pathway figures that reflect cancer-specific metabolic changes⁶⁹.

MetaboLights is one of the most important experimental repositories in the metabolomics layer because it stores metabolomics studies and associated experimental data. Its key significance lies in reusability and data access. Just as GEO and SRA enable reanalysis in transcriptomics, MetaboLights provides a public platform through which investigators can identify and reuse metabolomics datasets generated by prior studies. In cancer data retrieval, this is highly valuable for validating reported metabolic signatures, comparing metabolite patterns across studies, and performing secondary analyses without generating new experimental data. MetaboLights therefore functions as a major source of real-world metabolomics evidence in computational oncology⁷⁰.

Small Molecule Pathway Database (SMPDB) has a distinct value because it focuses specifically on human small-molecule pathways. Its importance lies in its disease-oriented and pathway-centered design, which makes it particularly helpful in translational interpretation. In cancer research, SMPDB is useful when altered metabolites need to be mapped onto human-specific biochemical

pathways with strong biomedical relevance. Compared with broader metabolic resources, SMPDB is especially suitable for studies that seek to connect tumor-associated metabolic changes to clinically interpretable human

pathways⁷¹. This gives it a particularly important place in the later interpretive stages of cancer metabolomics. The principal metabolomics databases used in cancer bioinformatics are summarized in Table 7.

Table 7. Metabolomics databases for in silico cancer data retrieval

Database	Description	Link
KEGG	Database of metabolic pathways, genes, and biochemical reactions	https://www.kegg.jp
HMDB (Human Metabolome Database)	Detailed information on human metabolites and their biological roles	https://hmdb.ca
Reactome	Curated database of biological pathways including metabolism and signaling	https://reactome.org
PubChem	Large chemical database containing compound structures and properties	https://pubchem.ncbi.nlm.nih.gov
PathWhiz	Tool for visualizing and constructing metabolic pathways	https://smpdb.ca/pathwhiz
MetaboLights	Repository for metabolomics experimental data and studies	https://www.ebi.ac.uk/metabolights
SMPDB	Database of small molecule pathways in humans	https://smpdb.ca

These metabolomics databases form a coherent and complementary framework for cancer data retrieval. KEGG and Reactome provide pathway-level structure; HMDB and PubChem support metabolite-level identity and characterization; PathWhiz enables pathway visualization; MetaboLights offers access to experimental metabolomics studies; and SMPDB strengthens human pathway interpretation. Used sequentially, these resources allow investigators to move from a metabolic signal or compound name toward a far more informative understanding of biochemical dysregulation in cancer. In the broader logic of multi-omics oncology, metabolomics databases are especially important because they widen the analytical lens from upstream molecular alteration to downstream biological consequence, thereby completing the functional interpretation of *in silico* cancer data retrieval. Figure 4 shows how metabolomics and pathway databases can be integrated to achieve downstream

biochemical interpretation in cancer bioinformatics.

Metabolomics Tools

Metabolomics tools are indispensable for processing, identifying, and interpreting metabolic features associated with cancer. MetaboAnalyst is a widely used platform for statistical analysis, visualization, and biological interpretation of metabolomics data, and it is particularly useful for identifying significant metabolic alterations and linking them to relevant biochemical pathways⁷². XCMS serves a critical preprocessing role by performing peak detection, alignment, and quantification of LC-MS data, thereby enabling the conversion of raw metabolomics signals into analyzable features⁷³. MetFrag contributes at the metabolite-identification stage by matching MS/MS spectra with candidate structures, which helps researchers assign chemical identity to detected metabolic signals. Together, these tools allow metabolomics datasets to progress from raw

instrumental output to biologically meaningful insights into metabolic dysregulation and pathway-level changes in cancer⁷⁴. The main

metabolomics analytical tools used for metabolite analysis and pathway interpretation are presented in Table 8.

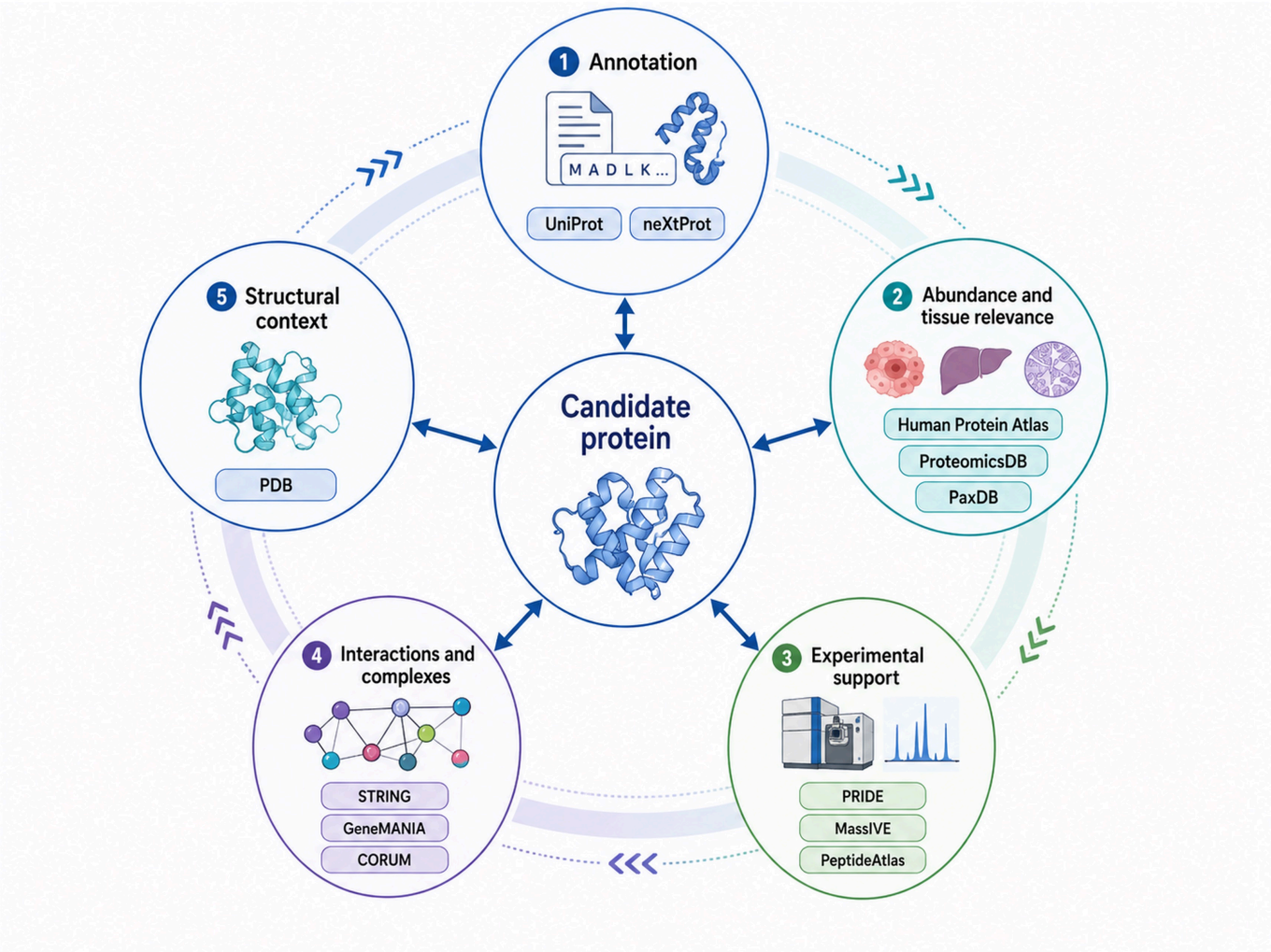


Figure 4. Metabolomics and pathway integration in cancer bioinformatics. Metabolite identification and experimental datasets are integrated with pathway databases and visualization tools to suppdysregulation of metabolic reprogramming and pathway dysregulation in cancer.

Table 8. Metabolomics tools for metabolite analysis and pathway interpretation

Tool	Description	Link
MetaboAnalyst	Platform for statistical analysis, visualization, and interpretation of metabolomics data	https://www.metaboanalyst.ca
XCMS	Processes LC-MS data for peak detection, alignment, and quantification	https://xcmsonline.scripps.edu
MetFrag	Identifies metabolites by matching MS/MS spectra with candidate structures	https://ipb-halle.github.io/MetFrag

INTERACTION DATABASES in CANCER DATA RETRIEVAL

Interaction databases occupy an essential position in *in silico* cancer research because malignant phenotypes rarely arise from isolated genes or proteins acting alone. Instead, tumor behavior is usually driven by

disturbed interaction networks, altered pathway connectivity, and disrupted multiprotein assemblies⁷⁵. For this reason, interaction-centered resources are critical after the initial identification of candidate genes, transcripts, or proteins, as they help

determine whether these candidates participate in biologically meaningful networks, functional modules, or regulatory complexes⁷⁶. Within the resource framework provided in the attached tables, the most clearly interaction-oriented platforms are GeneMANIA, STRING, and CORUM. Each contributes a distinct layer of network interpretation and should be used according to its specific strength rather than treated as interchangeable.

GeneMANIA is particularly valuable at the gene-centered stage of cancer data retrieval. It is designed to identify genes that are related to a user-supplied gene or gene set by integrating multiple forms of functional association data, including protein and genetic interactions, pathways, co-expression, co-localization, and protein domain similarity. Its importance lies in its predictive and hypothesis-generating role. In cancer studies, once a candidate gene has been identified through genomic or transcriptomic analysis, GeneMANIA can be used to determine whether that gene is functionally linked to known oncogenic pathways, whether it clusters with genes of similar biological function, or whether it may represent an overlooked member of a cancer-relevant pathway or protein complex. Thus, GeneMANIA is especially useful in the early interpretive phase, where the investigator is expanding from a single-gene finding toward a broader gene-network context⁷⁷.

STRING is one of the most important resources for protein-centered interaction analysis in cancer bioinformatics. It is a database of known and predicted protein–protein interactions that includes both direct physical interactions and indirect functional associations. These associations are derived

from multiple evidence channels, including high-throughput experiments, curated databases, co-expression, text mining, and genomic context prediction. The importance of STRING in cancer data retrieval lies in its ability to transform a list of candidate genes or proteins into an interpretable interaction network. In practical terms, this means that a protein identified as differentially expressed or mutation-associated can be examined in relation to its partners, pathway neighbors, and potential signaling modules. This is highly relevant in oncology, where network position may help distinguish a peripheral molecular change from a central regulatory hub. STRING is therefore particularly well suited to the intermediate-to-late stage of *in silico* analysis, when researchers need to evaluate whether candidate molecules converge within meaningful oncogenic circuits⁷⁸.

CORUM contributes a more specialized but highly important type of interaction evidence by focusing on experimentally characterized mammalian protein complexes. Unlike broader interaction databases that may include predicted or indirect associations, CORUM is centered on curated information about multiprotein complexes and their biological properties. Its importance in cancer data retrieval lies in the fact that many key oncogenic processes are executed not by single proteins, but by stable or semi-stable complexes involved in transcriptional control, signal transduction, DNA repair, chromatin remodeling, and cell-cycle regulation. In this context, CORUM helps determine whether a candidate protein should be interpreted as part of a known functional complex and whether disruption of that complex may be mechanistically relevant to tumor biology. As a result, CORUM is particularly valuable in the later interpretive stage of cancer

bioinformatics, when the objective is to move from network association toward more concrete mechanistic explanation at the level of molecular assemblies⁷⁹.

In cancer research, these interaction databases provide a logical hierarchy of network interpretation. GeneMANIA is especially useful for gene-level functional association and candidate expansion, STRING is most effective for protein-level interaction mapping and network analysis, and CORUM is most informative for experimentally validated mammalian protein complexes and mechanistic complex-level interpretation. When used sequentially, they provide a strong analytical bridge between primary omics findings and systems-level biological understanding. In a robust *in silico* cancer workflow, interaction databases should therefore be regarded not as optional add-ons, but as essential interpretive resources that help convert molecular candidates into biologically credible cancer mechanisms.

PATHWAY DATABASES in CANCER DATA RETRIEVAL

Pathway databases are indispensable in *in silico* cancer research because they allow investigators to move beyond lists of altered genes, transcripts, proteins, or metabolites and interpret those changes within organized biological systems⁸⁰. In cancer, isolated molecular abnormalities rarely act alone; rather, they converge on signaling cascades, metabolic circuits, biosynthetic routes, and regulatory modules that collectively shape tumor initiation, progression, metastasis, immune evasion, and therapeutic resistance. For this reason, pathway databases play a central interpretive role in cancer data retrieval⁸¹. They help determine whether a candidate molecule belongs to a known

biological pathway, whether several altered molecules converge on the same process, and whether apparently unrelated omics findings can be integrated into a coherent mechanistic framework⁸². Within your attached resource inventory, the principal pathway-oriented databases are KEGG, Reactome, PathWhiz, and SMPDB, with HMDB and PubChem also contributing important pathway-supportive functions through metabolite and compound annotation.

Kyoto Encyclopedia of Genes and Genomes is one of the most widely used pathway databases in biomedical research and is especially important in cancer bioinformatics because it integrates genes, enzymes, metabolites, and biochemical reactions into unified pathway maps. Its strength lies in providing a structured systems-level view in which molecular alterations can be placed into recognizable biological processes such as glycolysis, the tricarboxylic acid cycle, amino acid metabolism, lipid synthesis, apoptosis, or growth-factor signaling. In cancer data retrieval, KEGG is particularly valuable when researchers want to determine whether a dysregulated gene or metabolite belongs to a known oncogenic or metabolic pathway. It is often the first pathway resource used after initial omics screening because it translates molecular findings into interpretable biological routes.

Reactome is equally important, but it differs from KEGG in the depth and style of pathway curation. Reactome is a curated pathway database that includes both metabolism and signaling, making it particularly useful in cancer studies where metabolic changes are tightly linked to regulatory and signaling networks. Its strength lies in high-quality pathway organization and

the ability to connect molecular findings to broader functional systems rather than metabolism alone. In practical terms, Reactome is especially useful when a cancer study aims to interpret altered genes or proteins in the context of coordinated signaling events, immune pathways, cell-cycle regulation, or apoptosis-related processes in addition to classical metabolism. Thus, Reactome is highly valuable for broad mechanistic interpretation and cross-pathway integration.

PathWhiz occupies a more specialized but highly useful position in pathway-based cancer analysis. Its main importance lies in pathway visualization and pathway construction. Whereas KEGG and Reactome are often used for pathway reference and mapping, PathWhiz is particularly valuable when investigators need to visualize metabolic relationships clearly or build publication-quality pathway representations from retrieved data. In cancer studies, this becomes especially relevant when altered metabolites, enzymes, or pathway nodes need to be assembled into a readable biological model. PathWhiz therefore contributes strongly to interpretation, communication, and figure generation, making it especially useful in integrative and review-oriented *in silico* oncology.

Small Molecule Pathway Database is important because it focuses specifically on small-molecule pathways in humans. Its relevance in cancer data retrieval lies in its human-centered pathway design and its close alignment with disease-related biochemical interpretation. In contrast to more general pathway resources, SMPDB is particularly well suited for studies that aim to understand how small molecules, metabolites, or drugs participate in human biochemical pathways. In

cancer research, this makes it especially useful for interpreting altered metabolite signatures, treatment-associated metabolic changes, and biochemical pathways that may have translational or clinical relevance. Thus, SMPDB provides an important disease-oriented complement to broader pathway databases.

Although HMDB (Human Metabolome Database) is primarily a metabolite database rather than a classical pathway database, it is highly important in pathway retrieval because pathways cannot be interpreted properly without reliable metabolite-level annotation. HMDB provides detailed information on human metabolites and their biological roles, thereby helping researchers understand the identity, function, and human relevance of molecules that appear in cancer metabolomics studies. In pathway-centered cancer analysis, HMDB is especially useful when the investigator needs to move from pathway maps to the detailed biological meaning of individual metabolites within those pathways. It therefore supports pathway interpretation by adding compound-specific biological depth.

Similarly, PubChem is not a pathway database in the strict sense, but it plays an important supportive role in pathway-oriented cancer retrieval because it provides compound structures, properties, and identifiers. This is valuable when pathway interpretation depends on precise chemical characterization of metabolites, drugs, or small molecules. In cancer research, PubChem helps ensure that pathway-linked compounds are correctly identified and chemically contextualized, which is especially important in metabolomics, drug-response studies, and compound-centered pathway exploration. Its contribution is therefore

indirect but still essential to accurate pathway analysis.

These resources form a layered pathway-interpretation framework for *in silico* cancer research. KEGG provides broad integrative pathway maps connecting genes, enzymes, and metabolites; Reactome offers deeply curated pathways spanning metabolism and signaling; PathWhiz strengthens visualization and custom pathway representation; SMPDB focuses on human small-molecule pathways with strong disease relevance; and HMDB and PubChem supply metabolite- and compound-level detail that supports accurate pathway annotation. Used collectively, these databases enable researchers to move from isolated molecular findings toward a systems-level understanding of cancer biology, which is one of the central goals of multi-omics data retrieval.

CLINICAL and TRANSLATIONAL and FUNCTIONAL RESOURCES in CANCER RESEARCH

Clinical and Translational Databases in Cancer Research

Clinical and translational cancer databases provide the essential link between molecular findings and real-world oncologic relevance. Unlike purely omics-focused repositories, these resources help connect cancer-associated genomic alterations to clinical outcomes, therapeutic implications, trial opportunities, and population-level disease patterns. Among the databases listed in the attached table, The Cancer Genome Atlas and the International Cancer Genome Consortium are especially important because they combine large-scale genomic and multi-omics datasets with clinical information across multiple cancer types and populations⁸³. OncoKB and CIViC provide a more directly translational layer by linking cancer-related variants to therapeutic significance, levels of

clinical evidence, and expert interpretation. cBioPortal is particularly useful for interactive exploration of genomic datasets alongside survival and clinical variables, making it highly valuable for integrative cancer analysis⁸⁴. ClinicalTrials (<https://clinicaltrials.gov/>) extends this framework into the therapeutic domain by allowing researchers and clinicians to identify ongoing and completed cancer trials, treatment strategies, and eligibility criteria⁸⁵. Finally, SEER provides a population-based epidemiologic perspective through incidence, survival, and demographic statistics, thereby strengthening the public health and outcome-oriented interpretation of cancer research⁸⁶. Together, these databases form a critical clinical evidence layer that complements molecular databases and supports translational oncology, precision medicine, and outcome-focused cancer research. The principal clinical and translational databases included in this review are summarized in the attached table. The major clinical and translational databases relevant to cancer research are presented in Table 9.

DepMap is an important functional resource in cancer research because it extends *in silico* analysis beyond descriptive molecular profiling toward the identification of cancer vulnerabilities. Unlike databases that primarily catalogue genomic alterations, expression patterns, or clinical associations, DepMap enables investigation of gene essentiality and dependency across large panels of cancer cell lines. This makes it particularly valuable for determining whether genes or pathways identified through genomics, transcriptomics, or proteomics are not only altered in tumors but also selectively required for tumor survival. In the context of *in silico* cancer research, DepMap adds a

functional layer that strengthens biomarker evaluation, mechanistic interpretation, and therapeutic target prioritization, thereby supporting a more complete framework for precision oncology ⁸⁶.

Table 9. Clinical and translational databases used in cancer research

Database	Description	Link
The Cancer Genome Atlas	Large-scale project integrating multi-omics data (DNA, RNA, epigenetics) with clinical outcomes across many cancer types.	https://portal.gdc.cancer.gov
International Cancer Genome Consortium	International initiative collecting genomic and clinical data from diverse cancer populations worldwide.	https://dcc.icgc.org
OncoKB	Curated database linking cancer mutations to targeted therapies and clinical evidence levels.	https://www.oncokb.org
CIViC	Open, expert-curated database providing clinical interpretation of cancer-related genetic variants.	https://civicdb.org
cBioPortal	Interactive platform for exploring cancer genomic datasets with clinical data and survival analysis tools.	https://www.cbioportal.org
ClinicalTrials.gov	Global registry of clinical studies, including cancer trials, treatments, and patient eligibility data.	https://clinicaltrials.gov
SEER	Population-based database providing cancer incidence, survival, and demographic statistics.	https://seer.cancer.gov

AI-BASED TOOLS in CANCER GENOMICS and CLINICAL INTERPRETATION

Artificial intelligence has become increasingly important in cancer research because it improves the speed, consistency, and clinical relevance of genomic and transcriptomic data interpretation. The tools listed in the attached table represent a growing class of AI-enabled platforms that support variant prioritization, clinical reporting, phenotype-guided interpretation, and advanced transcriptomic modeling ⁸⁷.

Table 10. AI-based tools used in cancer genomics and clinical interpretation

Tool	Type	Description	Link
SeqOne	AI Clinical Genomics Platform	AI-powered platform for NGS analysis, variant prioritization, and clinical reporting for rare diseases and cancer genomics.	https://www.seqone.com
Franklin by Genoox	AI Variant Interpretation Tool	Cloud-based AI system that classifies and prioritizes genetic variants using ACMG guidelines and clinical evidence.	https://franklin.genoox.com
Fabric Genomics	AI Clinical Decision Platform	Uses AI and phenotype-driven analysis to interpret genomic data for rare diseases and precision medicine.	https://fabricgenomics.com
Genomic OS	AI Genomics Platform	Integrated AI system for genomic data analysis, variant interpretation, and clinical workflow support.	https://www.genomicos.com
VarSome	AI Variant Interpretation Platform	Combines multiple genomic databases with AI-based ACMG classification for variant annotation and interpretation.	https://varsome.com
scVI	AI Deep Learning Tool	Deep generative model for single-cell RNA-seq analysis, including clustering, batch correction, and visualization.	https://scvi-tools.org

In the clinical genomics setting, SeqOne, Franklin by Genoox, Fabric Genomics,

GenomicOS, and VarSome are particularly valuable because they integrate genomic data analysis with AI-assisted interpretation frameworks⁸⁸. These platforms are designed to support next-generation sequencing analysis, variant annotation, ACMG-based classification, phenotype-driven prioritization, and clinical decision support, which are especially important in precision oncology and rare-disease diagnostics. In parallel, scVI represents a different but equally important type of AI tool, as it applies deep learning to single-cell RNA sequencing data for clustering, batch correction, and visualization. This makes it highly relevant for cancer studies involving tumor heterogeneity, cell-state analysis, and transcriptomic integration⁸⁹. Collectively, these AI-driven tools extend conventional bioinformatics by enabling more automated, scalable, and clinically oriented interpretation of complex cancer datasets. The AI-related tools included in this review are summarized in the attached table. The principal AI-based tools used in cancer genomics and clinical interpretation are summarized in Table 10.

ANALYTICAL TOOLS USED in CANCER DATA RETRIEVAL

The analytical value of cancer databases depends fundamentally on the tools used to process, compare, validate, and interpret the retrieved data. Public repositories provide the raw material of *in silico* oncology, but analytical tools transform that material into biologically meaningful evidence⁹⁰. In practice, these tools occupy the operational center of cancer bioinformatics: they align sequencing reads, assemble transcripts, identify variants, quantify proteins, detect metabolites, reconstruct interaction networks, and support downstream statistical interpretation⁹¹. For this reason, analytical tools should not be

treated as auxiliary software appended after database retrieval; rather, they are the computational engines through which retrieved datasets become interpretable. In the context of the attached resource framework, the analytical tools can be understood in terms of the omics layer in which they operate and the specific type of evidence they generate⁹².

Within the genomics layer, analytical tools are primarily concerned with sequence comparison, read alignment, variant identification, mutational interpretation, visualization, and network-based biological inference. GeneMANIA is particularly valuable when the investigator begins with one or more candidate genes and seeks to expand those findings into a functionally relevant gene network. Its importance lies in gene-function prediction and gene–gene interaction inference, allowing researchers to determine whether newly identified candidates cluster with established cancer-associated genes or pathways. In this way, GeneMANIA helps move from isolated genomic signals to broader biological interpretation⁷⁷. BLAST serves a different but equally important purpose by identifying sequence similarity between DNA or protein sequences. In cancer studies, BLAST is especially useful when sequence identity, homology confirmation, or reference comparison is needed before more specialized downstream analyses are performed⁹³.

MutaGene contributes a more cancer-specific interpretive function within genomics. Its importance lies in the analysis of somatic mutational signatures and in the distinction between likely driver and passenger mutations in tumor genomes. Because many observed somatic mutations in cancer are biologically

incidental, tools such as MutaGene are especially valuable for prioritizing variants that may carry stronger mechanistic or oncogenic relevance. BWA and Bowtie2 are foundational alignment tools in genomic workflows. Their shared importance lies in aligning short DNA sequencing reads to a reference genome, which is an essential preliminary step before variant calling can occur. Although they serve similar general purposes, their use often depends on the design of the sequencing workflow and the computational preferences of the analyst. GATK then takes on a central role in variant discovery, especially for SNPs and indels, and is therefore one of the most important tools in cancer genomics when moving from aligned sequence data to high-confidence variant calls. Finally, IGV provides an indispensable visualization layer, allowing researchers to inspect alignments and variants directly. This is especially important in cancer genomics because computational calls often require visual confirmation to rule out artifacts and to interpret locus-specific sequence behavior with confidence.

Within the transcriptomics layer, analytical tools are primarily used to convert raw or semi-processed RNA data into biologically meaningful expression information. STAR is one of the most important aligners in this domain because it is optimized for RNA-seq reads and supports splice-junction detection, which is critical for transcriptomic analysis. Its importance lies in accurately aligning RNA-derived reads to the genome while accounting for exon-intron structure and splicing complexity⁴⁴. StringTie follows this stage by assembling transcripts and estimating their abundance from RNA-seq data. In practical cancer studies, this is highly valuable because transcript-level abundance, isoform diversity, and expression estimation

often form the basis of downstream tumor-normal comparison⁴⁵.

The statistical core of transcriptomics is represented by Limma, edgeR, and DESeq2, each of which plays a major role in differential-expression analysis. Limma is especially well established for microarray analysis and has also been adapted for RNA-seq workflows, making it valuable in studies involving mixed expression platforms or reanalysis of legacy cancer datasets. edgeR is important because it is specifically designed for count-based RNA-seq data and provides robust statistical modeling for differential gene-expression analysis. DESeq2 serves a similar central role and is particularly valued for its negative binomial framework and stable performance in identifying differentially expressed genes from count data. In cancer transcriptomics, these tools are often decisive because they determine which genes, transcripts, or regulatory RNAs emerge as statistically significant candidates for further biological interpretation. Thus, while databases provide the expression data, these analytical tools determine how confidently those data can be translated into cancer-relevant expression signatures⁹⁴.

Within the proteomics layer, analytical tools are primarily focused on protein identification, quantification, and targeted validation from mass spectrometry data. MaxQuant is one of the most important tools in this group because it supports large-scale protein identification and quantification from mass spectrometry experiments. Its value in cancer studies lies in its ability to turn raw proteomics data into interpretable protein abundance profiles, which are essential when validating whether transcriptomic or genomic signals are reflected at the protein level⁹⁵.

Skyline occupies a complementary role and is especially important in targeted proteomics and quantitative mass spectrometry analysis. In practical terms, Skyline is useful when investigators need more focused validation of specific peptides or proteins, particularly in candidate-driven biomarker studies⁹⁶. ProteinPilot also supports protein identification from MS/MS data and contributes an additional route for translating raw spectrometric evidence into protein-level biological interpretation. These tools enable researchers to move from proteomics repositories to experimentally grounded evidence about protein presence, abundance, and relevance in cancer.

Within the metabolomics layer, the principal analytical tasks include peak detection, alignment, metabolite identification, statistical interpretation, and pathway-level contextualization. MetaboAnalyst is one of the most influential tools in this domain because it provides a broad platform for statistical analysis, visualization, and biological interpretation of metabolomics data. Its importance lies in the fact that metabolomics datasets are often highly multidimensional and require integrated statistical and pathway-oriented exploration. In cancer research, MetaboAnalyst is especially valuable for identifying significant metabolic shifts, comparing tumor-associated metabolite patterns, and linking those changes to biological pathways⁹⁷. XCMS serves a more preprocessing-oriented function by handling LC-MS peak detection, alignment, and quantification. This makes it foundational when raw metabolomics data must first be converted into analyzable metabolic features⁹⁸. MetFrag then contributes at the identification stage by matching MS/MS

spectra with candidate structures, thereby helping determine which metabolites are actually represented by observed spectral signals. In combination, these three tools provide the computational backbone for turning raw metabolomics data into interpretable evidence about altered biochemical states in cancer⁷⁴.

When considered together, these analytical tools form a multi-omics computational infrastructure for cancer data retrieval. Genomics tools identify and validate DNA-level changes; transcriptomics tools quantify and compare expression; proteomics tools detect and measure protein-level evidence; and metabolomics tools reveal biochemical consequence. Their collective importance lies in the fact that they do not merely analyze data independently within each omics layer—they also enable cross-layer interpretation. A variant detected through genomic tools may be followed by transcriptomic quantification, proteomic confirmation, and metabolomic consequence mapping, thereby transforming a single molecular observation into a coherent, systems-level cancer hypothesis. For this reason, analytical tools should be viewed as a central interpretive framework in *in silico* oncology, linking database retrieval to biological meaning and translational relevance.

Cytoscape is a widely used open-source platform for the visualization and analysis of molecular interaction networks and biological pathways. Its importance in *in silico* cancer research lies in its ability to integrate omics-derived measurements, such as gene expression, protein interactions, and pathway annotations, into interpretable network models. This makes Cytoscape especially valuable after candidate genes, proteins, or

pathways have been identified, as it allows researchers to explore network topology, visualize pathway relationships, and generate publication-ready interaction maps. In this way, Cytoscape serves as an important cross-omics analytical and visualization tool that strengthens systems-level interpretation in cancer bioinformatics⁹⁹.

DATA-MINING LOGIC: from RESOURCE DISCOVERY to CROSS-OMICS EVIDENCE

The logic of data mining in cancer bioinformatics begins not with downloading datasets, but with defining the biological question precisely enough to determine which class of resource is required. A mutation-centered question demands a different entry point from an expression-centered, protein-centered, or pathway-centered question. Once the question has been clarified, the first operational step is usually target framing¹⁰⁰. In this stage, *GeneCards* is particularly valuable because it provides a gene-centered overview of function, disease association, and biological context, allowing the investigator to move from a candidate name to a structured molecular profile. *Ensembl* then becomes important for genome annotation, transcript structure, and comparative genomic context, helping the researcher define the exact genomic entity under investigation. *GenBank* complements this step by providing direct access to nucleotide sequences and reference records, which is essential when sequence confirmation or accession-based retrieval is required¹⁰¹.

Once the gene or locus has been defined, the next step is background filtering, where the goal is to distinguish potentially important alterations from ordinary human variation. gnomAD is central at this stage because it aggregates large-scale human genetic

variation and allows allele-frequency assessment across populations. The 1000 Genomes Project provides an additional population reference framework and is especially useful for understanding the distribution of common human variants across diverse ancestries¹⁰². dbSNP adds a variant-centered reference layer by cataloging SNPs and other small genetic changes and helping standardize variant identity. In data-mining logic, these three databases collectively prevent the premature interpretation of common background polymorphisms as cancer-relevant findings. They function as the filtering gate between raw variation and biologically plausible disease association.

After background filtering, the workflow shifts to cancer-specific genomic interpretation. COSMIC is a key resource at this point because it catalogs somatic mutations observed in human cancers and therefore allows researchers to determine whether a candidate alteration has already been reported in tumor contexts. The GDC Portal extends this interpretation by providing access to cancer genomic datasets linked to clinical information, which makes it particularly important when the investigator wants to connect genomic findings with tumor type, patient features, or broader molecular context. ClinVar becomes highly relevant when a variant may have known clinical significance, while dbVar is essential when the candidate event involves larger-scale structural variation such as deletions or duplications rather than only small sequence changes. The UCSC Genome Browser then provides a crucial visualization and contextualization layer, allowing the locus to be explored within the annotated genome. In a professional cancer data-mining workflow, these resources

collectively convert a filtered genomic signal into a cancer-informed and clinically interpretable genomic observation.

The next major stage is transcript-level validation, where the investigator asks whether the genomic signal is reflected in RNA expression. GEO is particularly important here because it provides public gene-expression studies from microarray and RNA-seq experiments and is highly useful for comparing expression patterns across independent cancer studies. TCGA serves as one of the principal tumor-centered transcriptomic resources and is especially important for large-scale cancer expression analysis linked to clinical and genomic annotations. GTEx is indispensable as the normal-reference counterpart because it provides healthy tissue expression profiles and thereby helps determine whether an apparent tumor signal is truly dysregulated rather than simply tissue specific. SRA supports raw-data reuse and becomes especially important when the investigator wishes to reanalyze RNA-seq data from original sequence files rather than depend on processed results. In this stage of the workflow, transcriptomic evidence is used to decide whether a genomic candidate is transcriptionally active, suppressed, or differentially regulated in cancer.

Transcript-level data mining frequently requires a more refined view of regulatory and cellular complexity. Single Cell Expression Atlas is important because it resolves expression at the level of individual cells and allows investigators to determine whether a signal arises from malignant cells, stromal components, or immune populations within the tumor microenvironment. NONCODE contributes a dedicated resource for non-coding RNA sequences and annotation, which

is critical in cancer studies involving long non-coding RNAs and other regulatory transcripts. miRBase serves the same purpose for microRNAs and is particularly important when post-transcriptional regulation is part of the biological hypothesis. Expression Atlas broadens the analysis by placing genes and transcripts into a wider condition- and tissue-based expression context, while FANTOM5 adds a regulatory dimension by incorporating promoter and enhancer RNA expression. BioGPS functions as a supportive expression and annotation portal that helps researchers obtain a broader overview of gene-expression distribution. Together, these databases deepen the transcriptomics phase by moving from bulk abundance toward regulatory, contextual, and cell-specific interpretation.

Once expression evidence has been established, the data-mining workflow moves to the protein level, where the question becomes whether transcriptional abnormalities are translated into functional protein relevance. UniProt is usually the first major resource at this stage because it provides comprehensive protein sequence and functional annotation, helping define the biological properties of the encoded protein. The Human Protein Atlas is particularly important because it places protein expression into tissue and cellular context, which is highly relevant in cancer where localization and tissue specificity matter. ProteomicsDB adds quantitative human proteome information, while neXtProt provides high-quality human protein knowledge with strong functional annotation. PaxDB strengthens this phase by offering abundance information across tissues and organisms, thereby helping contextualize whether a protein is generally abundant or unusually represented in a cancer setting. These

databases together determine whether the transcript-level candidate is biologically embodied as a protein signal and whether that signal is credible within human tissue biology.

Protein-level data mining is then strengthened by repositories and interaction resources that connect candidates to experimental and systems evidence. PRIDE Archive is essential because it stores public mass spectrometry-based proteomics data and therefore allows reuse of experimental datasets for protein confirmation. MassIVE serves a similar large-scale repository role and is valuable when broader proteomics reanalysis is needed. PeptideAtlas is especially useful at the peptide-evidence level because it records identified peptides and proteins from mass spectrometry experiments. STRING becomes highly important when the goal shifts from protein presence to protein interaction, since it provides protein–protein interaction networks and functional associations that help determine whether a candidate belongs to a known oncogenic module. PDB adds structural depth by supplying three-dimensional protein data, which is critical when the mechanistic relevance of a mutation or protein change depends on structural consequence. CORUM completes this step by focusing on experimentally verified protein complexes, allowing investigators to determine whether a candidate protein should be interpreted as part of a larger molecular assembly. Together, these databases move the workflow from individual protein evidence toward functional network and complex-level interpretation.

The final widening stage of cross-omics evidence building concerns pathways and metabolic consequence. KEGG is especially important because it connects genes,

enzymes, metabolites, and biochemical reactions into integrated pathway maps, allowing molecular changes to be interpreted in a systems framework. Reactome adds a curated pathway perspective that includes both metabolism and signaling, making it useful when cancer-associated findings must be integrated across broader biological processes. HMDB contributes metabolite-centered biological interpretation and helps define the human relevance of specific metabolic compounds. PubChem provides detailed compound structures and chemical properties, which strengthens metabolite identification and annotation. PathWhiz is particularly valuable for pathway visualization and pathway construction, making it useful when altered molecules need to be assembled into biologically interpretable pathway models. MetaboLights provides access to experimental metabolomics studies and therefore supports secondary analysis of metabolic cancer datasets. SMPDB focuses on small-molecule pathways in humans and is especially useful when the investigator wants disease-oriented metabolic interpretation. At this final stage, data mining reaches its broadest analytical scope: candidate alterations are no longer interpreted only as isolated molecular events, but as components of coordinated biochemical and signaling networks that shape tumor biology.

Taken as a whole, the data-mining pathway from resource discovery to cross-omics evidence follows a progressive logic of increasing interpretive depth. Gene-centered resources define the target, population databases filter background variation, cancer-specific genomic resources establish tumor relevance, transcriptomic repositories test functional expression, proteomic databases confirm biological embodiment, and

pathway/metabolomics databases reveal systems-level consequence. The importance of each database lies in its specific role within this sequence. Strong *in silico* cancer research is therefore not produced by consulting many databases indiscriminately, but by using each

resource at the correct stage of evidence building and integrating the outputs into a coherent multi-omics interpretation. Figure 5 summarizes the stepwise workflow of *in silico* cancer data mining using public databases.

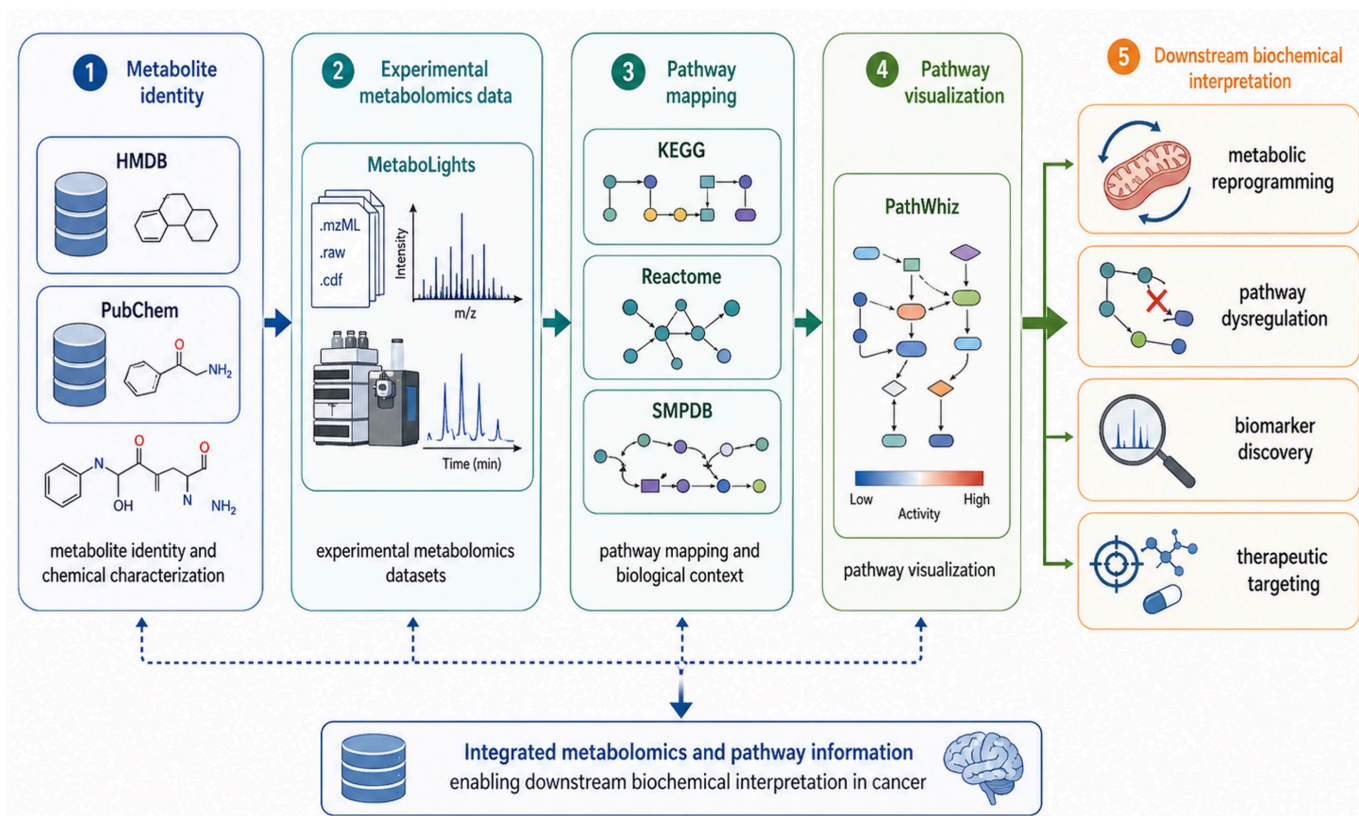


Figure 5. Stepwise workflow for cancer data mining using public databases. The workflow begins with research-question formulation and database discovery, followed by genomic, transcriptomic, proteomic, and metabolomic analysis, and concludes with cross-omics integration and biological interpretation.

CONCLUSIONS

Public multi-omics databases have become essential to *in silico* cancer research because they enable a structured transition from molecular data retrieval to biological and translational interpretation. This review shows that the real strength of computational oncology lies in the combined use of complementary resources across genomics, transcriptomics, proteomics, metabolomics, pathway analysis, interaction biology, and clinical evidence. When used sequentially, these databases support a coherent evidence-building workflow for biomarker discovery, mechanistic insight, and target prioritization. A key conclusion is that database discovery

should be treated as a formal methodological step, since different repositories serve distinct analytical purposes. In addition, findings from a single omics layer are rarely sufficient for strong inference; more reliable conclusions emerge when evidence is cross-validated across multiple molecular levels and interpreted within biological and clinical context. Future studies should therefore report database selection clearly, prioritize cross-omics validation, and adopt reproducible workflows that connect data retrieval, analytical processing, and integrative interpretation. Such an approach will strengthen the rigor, biological relevance, and translational value of *in silico* cancer research.

DECLARATIONS

Data Availability Statement

No new data were generated or analyzed in this review article.

Funding

The authors declare that no funding was received for this work.

Ethics Approval and Consent to Participate

Not applicable.

References

- Chakraborty S, Sharma G, Karmakar S, Banerjee S. Multi-OMICS approaches in cancer biology: New era in cancer therapy. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2024;1870(5):167120. doi: <https://doi.org/10.1016/j.bbadis.2024.167120>
- Zhang X, Xie J, Yang Z, Yu CKW, Hu Y, Qin J. Tumour heterogeneity and personalized treatment screening based on single-cell transcriptomics. *Computational and Structural Biotechnology Journal*. 2025;27:307-320. doi: <https://doi.org/10.1016/j.csbj.2024.12.020>
- Vitorino R. Transforming Clinical Research: The Power of High-Throughput Omics Integration. *Proteomes*. 2024;12(3):25. doi: <https://doi.org/10.3390/proteomes12030025>
- Grigoriadis D, Tsifintaris M, Giannakakis A, Pavlopoulos GA, Perdikopanis N. Public Omics Explorer (POE): Enabling integrative semantic search across GEO omics datasets based on PubMed publications. *Computational and Structural Biotechnology Journal*. 2025;27:4802-4812. doi: <https://doi.org/10.1016/j.csbj.2025.11.004>
- Putrama IM, Martinek P. Heterogeneous data integration: Challenges and opportunities. *Data in Brief*. 2024;56:110853. doi: <https://doi.org/10.1016/j.dib.2024.110853>
- Wolde T, Bhardwaj V, Pandey V. Current Bioinformatics Tools in Precision Oncology. *MedComm*. 2025;6(7):e70243. doi: <https://doi.org/10.1002/mco2.70243>
- Ma L, Zou D, Liu L, et al. Database Commons: A Catalog of Worldwide Biological Databases. *Genomics, Proteomics & Bioinformatics*. 2022;21(5):1054-1058. doi: <https://doi.org/10.1016/j.gpb.2022.12.004>
- Pacini C, Duncan E, Gonçalves E, et al. A comprehensive clinically informed map of dependencies in cancer cells and framework for target prioritization. *Cancer Cell*. 2024;42(2):301-316.e9. doi: <https://doi.org/10.1016/j.ccell.2023.12.016>
- Laskar FS, Bappy MNI, Hossain MS, et al. An In silico Approach towards Finding the Cancer-Causing Mutations in Human MET Gene. *International Journal of Genomics*. 2023;2023:1-19. doi: <https://doi.org/10.1155/2023/9705159>
- Antoszewski K, Chmielewska K, Jagiello K, Puzyn T. Designing RNA sequencing experiments: A practical guide to reproducible gene expression analysis. *Computational and Structural Biotechnology Journal*. 2025;31:101-119. doi: <https://doi.org/10.1016/j.csbj.2025.12.015>
- Jiang Z, Zhang H, Gao Y, Sun Y. Multi-omics strategies for biomarker discovery and application in personalized oncology. *Molecular Biomedicine*. 2025;6(1):115. doi: <https://doi.org/10.1186/s43556-025-00340-0>
- Mosha NF, Ngulube P. Metadata Standard for Continuous Preservation, Discovery, and Reuse of Research Data in Repositories by Higher Education Institutions: A Systematic Review. *Information*. 2023;14(8):427. doi: <https://doi.org/10.3390/info14080427>
- Ahmad B, Ouahada K, Hamam H. Machine learning for drug-target interaction prediction: A comprehensive review of models, challenges, and computational strategies. *Computational and Structural Biotechnology Journal*. 2026;31:316-345. doi: <https://doi.org/10.1016/j.csbj.2025.12.033>
- Molefi T, Mabonga L, Hull R, Sebitloane M, Dlamini Z. From Genes to Clinical Practice: Exploring the Genomic Underpinnings of Endometrial Cancer. *Cancers*. 2025;17(2):320. doi: <https://doi.org/10.3390/cancers17020320>
- Bonfiglio F, Legati A, Lasorsa VA, et al. Best practices for germline variant and DNA methylation analysis of second- and third-generation sequencing data. *Human Genomics*. 2024;18(1):120. doi: <https://doi.org/10.1186/s40246-024-00684-8>
- Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Research*. 2012;41(D1):D36-D42. doi: <https://doi.org/10.1093/nar/gks1195>
- Martin FJ, Amode MR, Aneja A, et al. Ensembl 2023. *Nucleic Acids Research*. 2022;51(D1):D933-D941. doi: <https://doi.org/10.1093/nar/gkac958>

Conflict of Interest

The authors declare no conflict of interest.

Use of Generative AI

Generative AI tools were used only for language improvement and proofreading. The authors reviewed and approved the final content and take full responsibility for the manuscript.

18. Gudmundsson S, Singer-Berk M, Watts NA, et al. Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation*. 2021;43(8):1012-1030. doi: <https://doi.org/10.1002/humu.24309>
19. Siva N. 1000 Genomes project. *Nature Biotechnology*. 2008;26(3):256-256. doi: <https://doi.org/10.1038/nbt0308-256b>
20. Phan L, Zhang H, Wang Q, et al. The evolution of dbSNP: 25 years of impact in genomic research. *Nucleic Acids Research*. 2024;53(D1):D925-D931. doi: <https://doi.org/10.1093/nar/gkae977>
21. Sondka Z, Dhir NB, Carvalho-Silva D, et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*. 2023;52(D1):D1210-D1217. doi: <https://doi.org/10.1093/nar/gkad986>
22. Wilson S, Fitzsimons M, Ferguson M, et al. Developing Cancer Informatics Applications and Tools Using the NCI Genomic Data Commons API. *Cancer Research*. 2017;77(21):e15-e18. doi: <https://doi.org/10.1158/0008-5472.can-17-0598>
23. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*. 2016;54(1):1-33. doi: <https://doi.org/10.1002/cpbi.5>
24. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*. 2015;44(D1):D862-D868. doi: <https://doi.org/10.1093/nar/gkv1222>
25. Lappalainen I, Lopez J, Skipper L, et al. dbVar and DGVA: public archives for genomic structural variation. *Nucleic Acids Research*. 2012;41(D1):D936-D941. doi: <https://doi.org/10.1093/nar/gks1213>
26. Perez G, Barber GP, Benet-Pages A, et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Research*. 2024;53(D1):D1243-D1249. doi: <https://doi.org/10.1093/nar/gkae974>
27. Franz M, Rodriguez H, Lopes C, et al. GeneMANIA update 2018. *Nucleic Acids Research*. 2018;46(W1):W60-W64. doi: <https://doi.org/10.1093/nar/gky311>
28. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 1997;25(17):3389-3402. doi: <https://doi.org/10.1093/nar/25.17.3389>
29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012;9(4):357-359. doi: <https://doi.org/10.1038/nmeth.1923>
30. Ren S, Bertels K, Al-Ars Z. Efficient Acceleration of the Pair-HMMs Forward Algorithm for GATK HaplotypeCaller on Graphics Processing Units. *Evolutionary Bioinformatics*. 2018;14:1176934318760543. doi: <https://doi.org/10.1177/1176934318760543>
31. Robinson JT, Thorvaldsdottir H, Turner D, Mesirov JP. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics*. 2022;39(1):btac830. doi: <https://doi.org/10.1093/bioinformatics/btac830>
32. Supplitt S, Karpinski P, Sasiadek M, Laczmanska I. Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *International Journal of Molecular Sciences*. 2021;22(3):1422. doi: <https://doi.org/10.3390/ijms22031422>
33. Austin BK, Firooz A, Valafar H, Blenda AV. An Updated Overview of Existing Cancer Databases and Identified Needs. *Biology*. 2023;12(8):1152. doi: <https://doi.org/10.3390/biology12081152>
34. Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. 2002;30(1):207-210. doi: <https://doi.org/10.1093/nar/30.1.207>
35. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking*. 2015;13(5):307-308. doi: <https://doi.org/10.1089/bio.2015.29031.hmm>
36. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Współczesna Onkologia*. 2015;1A(1a):68-77. doi: <https://doi.org/10.5114/wo.2014.47136>
37. Park SJ, Yoon BH, Kim SK, Kim SY. GENT2: an updated gene expression database for normal and tumor tissues. *BMC Medical Genomics*. 2019;12(S5):101. doi: <https://doi.org/10.1186/s12920-019-0514-7>
38. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Research*. 2010;39(Database):D19-D21. doi: <https://doi.org/10.1093/nar/gkq1019>
39. George N, Fexova S, Fuentes AM, et al. Expression Atlas update: insights from sequencing data at both bulk and single cell level. *Nucleic Acids Research*. 2023;52(D1):D107-D114. doi: <https://doi.org/10.1093/nar/gkad1021>
40. Xiyuan L, Dechao B, Liang S, et al. Using the NONCODE Database Resource. *Current Protocols in Bioinformatics*. 2017;58(1):1-19. doi: <https://doi.org/10.1002/cpbi.25>
41. Griffiths-Jones S. miRBase: microRNA Sequences and Annotation. *Current Protocols in Bioinformatics*. 2010;29(1):1-10. doi: <https://doi.org/10.1002/0471250953.bi1209s29>

42. Bertin N, Mendez M, Hasegawa A, et al. Linking FANTOM5 CAGE peaks to annotations with CAGEScan. *Scientific Data*. 2017;4(1):170147. doi: <https://doi.org/10.1038/sdata.2017.147>
43. Wu C, Jin X, Tsueng G, Afrasiabi C, Su AI. BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic Acids Research*. 2015;44(D1):D313-D316. doi: <https://doi.org/10.1093/nar/gkv1104>
44. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2012;29(1):15-21. doi: <https://doi.org/10.1093/bioinformatics/bts635>
45. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*. 2015;33(3):290-295. doi: <https://doi.org/10.1038/nbt.3122>
46. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47-e47. doi: <https://doi.org/10.1093/nar/gkv007>
47. Robinson MD, McCarthy DJ, Smyth GK. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139-140. doi: <https://doi.org/10.1093/bioinformatics/btp616>
48. Mani DR, Krug K, Zhang B, et al. Cancer proteogenomics: current impact and future prospects. *Nature Reviews Cancer*. 2022;22(5):298-313. doi: <https://doi.org/10.1038/s41568-022-00446-5>
49. Zhou Z, Zhang R, Zhou A, et al. Proteomics appending a complementary dimension to precision oncology. *Computational and Structural Biotechnology Journal*. 2024;23:1725-1739. doi: <https://doi.org/10.1016/j.csbj.2024.04.044>
50. The UniProt Consortium, Bateman A, Martin MJ, et al. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. 2022;51(D1):D523-D531. doi: <https://doi.org/10.1093/nar/gkac1052>
51. Persson A, Hober S, Uhlén M. A human protein atlas based on antibody proteomics. *Current Opinion in Molecular Therapeutics*. 2006;8(3):185-90.
52. Reisinger F, del-Toro N, Ternent T, Hermjakob H, Vizcaíno JA. Introducing the PRIDE Archive RESTful web services. *Nucleic Acids Research*. 2015;43(W1):W599-W604. doi: <https://doi.org/10.1093/nar/gkv382>
53. Desiere F. The PeptideAtlas project. *Nucleic Acids Research*. 2006;34(90001):D655-D658. doi: <https://doi.org/10.1093/nar/gkj040>
54. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *Journal of Proteome Research*. 2018;18(2):623-632. doi: <https://doi.org/10.1021/acs.jproteome.8b00702>
55. Noguchi T, Onizuka K, Akiyama Y, Saito M. PDB-REPRDB: a database of representative protein chains in PDB (Protein Data Bank). *Proceedings. International Conference on Intelligent Systems for Molecular Biology*. 1997;5:214-7.
56. Schmidt T, Samaras P, Frejno M, et al. ProteomicsDB. *Nucleic Acids Research*. 2017;46(D1):D1271-D1281. doi: <https://doi.org/10.1093/nar/gkx1029>
57. Gaudet P, Michel PA, Zahn-Zabal M, et al. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Research*. 2016;45(D1):D177-D182. doi: <https://doi.org/10.1093/nar/gkw1062>
58. Wang M, Weiss M, Simonovic M, et al. PaxDb, a Database of Protein Abundance Averages Across All Three Domains of Life. *Molecular & Cellular Proteomics*. 2012;11(8):492-500. doi: <https://doi.org/10.1074/mcp.o111.014704>
59. Tsitsiridis G, Steinkamp R, Giurgiu M, et al. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research*. 2022;51(D1):D539-D545. doi: <https://doi.org/10.1093/nar/gkac1015>
60. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*. 2016;11(12):2301-2319. doi: <https://doi.org/10.1038/nprot.2016.136>
61. Pino LK, Searle BC, Bollinger JG, Nunn B, MacLean B, MacCoss MJ. The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrometry Reviews*. 2017;39(3):229-244. doi: <https://doi.org/10.1002/mas.21540>
62. Jagtap P, Bandhakavi S, Higgins L, et al. Workflow for analysis of high mass accuracy salivary data set using MaxQuant and ProteinPilot search algorithm. *PROTEOMICS*. 2012;12(11):1726-1730. doi: <https://doi.org/10.1002/pmic.201100097>
63. Cai M, Liu H, Shao C, et al. Metabolomics and metabolites in cancer diagnosis and treatment. *Molecular Biomedicine*. 2025;6(1):109. doi: <https://doi.org/10.1186/s43556-025-00362-8>
64. Wishart D, Mandal R, Stanislaus A, Ramirez-Gaona M. Cancer Metabolomics and the Human Metabolome Database. *Metabolites*. 2016;6(1):10. doi: <https://doi.org/10.3390/metabo6010010>
65. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*. 2000;28(1):27-30. doi: <https://doi.org/10.1093/nar/28.1.27>
66. Wishart DS, Guo A, Oler E, et al. HMDB 5.0: the Human

- Metabolome Database for 2022. *Nucleic Acids Research*. 2021;50(D1):D622-D631. doi: <https://doi.org/10.1093/nar/gkab1062>
67. Milacic M, Beavers D, Conley P, et al. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*. 2023;52(D1):D672-D678. doi: <https://doi.org/10.1093/nar/gkad1025>
 68. Kim S, Chen J, Cheng T, et al. PubChem 2023 update. *Nucleic Acids Research*. 2022;51(D1):D1373-D1380. doi: <https://doi.org/10.1093/nar/gkac956>
 69. Pon A, Jewison T, Su Y, et al. Pathways with PathWhiz. *Nucleic Acids Research*. 2015;43(W1):W552-W559. doi: <https://doi.org/10.1093/nar/gkv399>
 70. Yurekten O, Payne T, Tejera N, et al. MetaboLights: open data repository for metabolomics. *Nucleic Acids Research*. 2023;52(D1):D640-D646. doi: <https://doi.org/10.1093/nar/gkad1045>
 71. Frolkis A, Knox C, Lim E, et al. SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Research*. 2009;38(suppl_1):D480-D487. doi: <https://doi.org/10.1093/nar/gkp1002>
 72. Chong J, Wishart DS, Xia J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Current Protocols in Bioinformatics*. 2019;68(1):e86. doi: <https://doi.org/10.1002/cpbi.86>
 73. Domingo-Almenara X, Siuzdak G. Metabolomics Data Processing Using XCMS. *Methods in Molecular Biology*. 2020;2104:11-24. doi: https://doi.org/10.1007/978-1-0716-0239-3_2
 74. Ruttkies C, Neumann S, Posch S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics*. 2019;20(1):376. doi: <https://doi.org/10.1186/s12859-019-2954-7>
 75. Schramm S, Li SS, Jayaswal V, et al. Disturbed protein-protein interaction networks in metastatic melanoma are associated with worse prognosis and increased functional mutation burden. *Pigment Cell & Melanoma Research*. 2013;26(5):708-722. doi: <https://doi.org/10.1111/pcmr.12126>
 76. De Silva MRP, Weeraman JWK, Piyatissa S, Fernando PC. Prediction of new candidate proteins and analysis of sub-modules and protein hubs associated with seed development in rice (*Oryza sativa*) using an ensemble network-based systems biology approach. *BMC Plant Biology*. 2025;25(1):604. doi: <https://doi.org/10.1186/s12870-025-06595-7>
 77. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*. 2008;9(S1):S4. doi: <https://doi.org/10.1186/gb-2008-9-s1-s4>
 78. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. 2018;47(D1):D607-D613. doi: <https://doi.org/10.1093/nar/gky1131>
 79. Tsitsiridis G, Steinkamp R, Giurgiu M, et al. CORUM: the comprehensive resource of mammalian protein complexes-2022. *Nucleic Acids Research*. 2022;51(D1):D539-D545. doi: <https://doi.org/10.1093/nar/gkac1015>
 80. Bagewadi ZK, Yaranguppi DA, Mulla SI, Patil VS, Naik MM. In-silico analysis, molecular docking and dynamic simulation of anti-cancer metabolites derived from *Bacillus velezensis*. *Medicine in Microecology*. 2026;27:100165. doi: <https://doi.org/10.1016/j.medmic.2026.100165>
 81. Sun YH, Zhang JX, Jin HS, Huang J. Crosstalk Between Metabolic Reprogramming and Epigenetic Modifications in Colorectal Cancer: Mechanisms and Clinical Applications. *Current Issues in Molecular Biology*. 2025;47(9):751. doi: <https://doi.org/10.3390/cimb47090751>
 82. Ryan V. WG, Sahay S, Vergis J, Weistuch C, Meller J, McCullumsmith RE. Pathway Analysis Interpretation in the Multi-Omic Era. *BioTech*. 2025;14(3):58. doi: <https://doi.org/10.3390/biotech14030058>
 83. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. 2013;45(10):1113-1120. doi: <https://doi.org/10.1038/ng.2764>
 84. Chakravarty D, Gao J, Phillips S, et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*. 2017;1:1-16. doi: <https://doi.org/10.1200/po.17.00011>
 85. Wu DT, Hanauer DA, Mei Q, et al. Assessing the readability of ClinicalTrials.gov. *Journal of the American Medical Informatics Association*. 2015;23(2):269-275. doi: <https://doi.org/10.1093/jamia/ocv062>
 86. Murphy PK, Sellers ME, Bonds SH, Scott S. The SEER Program's longstanding commitment to making cancer resources available. *JNCI Monographs*. 2024;2024(65):118-122. doi: <https://doi.org/10.1093/jncimonographs/lgae028>
 87. Liu K, Luo Z, Zhang W, et al. Artificial intelligence models: transforming early diagnosis and precise treatment of gastrointestinal cancers. *Molecular Cancer*. 2026;25(1):90. doi: <https://doi.org/10.1186/s12943-026-02589-7>
 88. Bastida-Lertxundi N, Martí-Carrera I, Laña-Ruiz B, et al. Evaluating seven bioinformatics platforms for tertiary analysis of genomic data from whole exome sequencing in a pilot group of patients. *Advances in Laboratory Medicine / Avances en Medicina de*

- Laboratorio. 2025;6(1):28-36. doi: <https://doi.org/10.1515/almed-2025-0031>
89. Ergen C, Pour Amiri VV, Kim M, et al. Scvi-hub: an actionable repository for model-driven single-cell analysis. *Nature Methods*. 2025;22(9):1836-1845. doi: <https://doi.org/10.1038/s41592-025-02799-9>
 90. Fawzy MP, Nafie MS, Nabih NW, et al. Reprogramming cancer therapeutics through artificial intelligence-guided drug repurposing: Computational innovation, experimental validation, and translational challenges. *Artificial Intelligence in the Life Sciences*. 2026;9:100168. doi: <https://doi.org/10.1016/j.ailsci.2026.100168>
 91. Zhang W, Chang JW, Lin L, et al. Network-Based Isoform Quantification with RNA-Seq Data for Cancer Transcriptome Analysis. *PLOS Computational Biology*. 2015;11(12):e1004465. doi: <https://doi.org/10.1371/journal.pcbi.1004465>
 92. Martínez-García M, Hernández-Lemus E. Data Integration Challenges for Machine Learning in Precision Medicine. *Frontiers in Medicine*. 2022;8:784455. doi: <https://doi.org/10.3389/fmed.2021.784455>
 93. Samal KC, Sahoo JP, Behera L, Dash T. Understanding the BLAST (Basic Local Alignment Search Tool) Program and a Step-by-step Guide for its use in Life Science Research. *Bhartiya Krishi Anusandhan Patrika*. 2021;36(Of):55-61. doi: <https://doi.org/10.18805/bkap283>
 94. Liu S, Wang Z, Zhu R, Wang F, Cheng Y, Liu Y. Three Differential Expression Analysis Methods for RNA Sequencing: limma, EdgeR, DESeq2. *Journal of Visualized Experiments*. 2021;175(175):e62528. doi: <https://doi.org/10.3791/62528>
 95. Pinter N, Glätzer D, Fahrner M, et al. MaxQuant and MSstats in Galaxy Enable Reproducible Cloud-Based Analysis of Quantitative Proteomics Experiments for Everyone. *Journal of Proteome Research*. 2022;21(6):1558-1565. doi: <https://doi.org/10.1021/acs.jproteome.2c00051>
 96. MacLean B, Tomazela DM, Shulman N, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*. 2010;26(7):966-968. doi: <https://doi.org/10.1093/bioinformatics/btq054>
 97. Pang Z, Lu Y, Zhou G, et al. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research*. 2024;52(W1):W398-W406. doi: <https://doi.org/10.1093/nar/gkae253>
 98. Louail P, Brunius C, Garcia-Aloy M, et al. xcms in Peak Form: Now Anchoring a Complete Metabolomics Data Preprocessing and Analysis Software Ecosystem. *Analytical Chemistry*. 2025;97(50):27639-27645. doi: <https://doi.org/10.1021/acs.analchem.5c04338>
 99. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*. 2003;13(11):2498-2504. doi: <https://doi.org/10.1101/gr.1239303>
 100. Sarkar MS, Mia MM, Amin MA, Hossain MS, Islam MZ. Bioinformatics and network biology approach to identifying type 2 diabetes genes and pathways that influence the progression of breast cancer. *Heliyon*. 2023;9(5):e16151. doi: <https://doi.org/10.1016/j.heliyon.2023.e16151>
 101. Sleptsova E, Vershinina O, Ivanchenko M, Turubanova V. From Data to Decision: Integrating Bioinformatics into Glioma Patient Stratification and Immunotherapy Selection. *International Journal of Molecular Sciences*. 2026;27(2):667. doi: <https://doi.org/10.3390/ijms27020667>
 102. Devuyst O. The 1000 Genomes Project: Welcome to a New World. *Peritoneal Dialysis International: Journal of the International Society for Peritoneal Dialysis*. 2015;35(7):676-677. doi: <https://doi.org/10.3747/pdi.2015.00261>